Systemic Modeling of Biomolecular Interaction Networks

by

Ali A. Sobhi Afshar

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

September, 2016

Abstract

MicroRNAs (miRNAs) are small non-coding ribonucleic acids (RNAs) that extensively regulate gene expression in metazoan animals, plants and protozoa. Approximately 22 nucleotides in length, miRNAs usually repress gene expression by binding to sequences with partial complementarity on target messenger RNA (mRNA) transcripts. In mammals, miRNAs are thought to control the activity of more than 60% of all protein-coding genes and extensively participate in the regulation of many cellular functions.

With few exceptions, metazoan miRNAs base-pair with their targets imperfectly, following a set of rules that have been formulated by employing experimental and bioinformatics-based analyses. This limited complementarity makes the task of computationally identifying miRNA targets very challenging and usually leads to large numbers of, mostly false, potential targets.

Earlier computational tools have mainly focused on dissecting individual miRNA-target interactions by relying on sequence-based identification of miRNA-target binding sites or on mRNA/miRNA expression data analysis. With the goal to gain a systemic understanding of miRNA-mediated interaction networks, in this research work, we develop IntegraMiR,

a novel integrative analysis method that can be used to infer certain types of regulatory loops of dysregulated miRNA/Transcription Factor (TF) interactions which appear at the transcriptional, post-transcriptional and signaling levels in a statistically over-represented manner.

To reliably predict miRNA-target interactions from mRNA/miRNA expression data and construct their networks with a systems perspective, our proposed method collectively utilizes the aforementioned molecular structures identified to be statistically over-represented in gene regulatory networks, sequence-based miRNA-target predictions obtained from different algorithms, known information about mRNA and miRNA targets of TFs available in existing databases, available molecular subtyping information, and state-of-the-art statistical techniques to appropriately constrain the underlying analysis.

To investigate the effectiveness of our proposed procedure, we apply our method on mRNA/miRNA expression data from prostate tumor and normal samples and detect numerous known and novel miRNA-mediated dysregulated loops and networks in prostate cancer. In addition, as an application of our work on miRNA/TF-mediated loop and network identification, we utilize the IntegraMiR paradigm and exploit additional publicly available databases and datasets to infer miRNA-mediated regulatory networks in Autism Spectrum Disorders, that were of interest to our experimental collaborators at the School of Medicine.

We demonstrate instances of the results in a number of distinct biological settings, which are known to play crucial roles in the contexts of prostate cancer and autism spec-

trum disorders. Our findings show that the proposed computational method can be used to effectively achieve notable systemic insights into the poorly understood molecular mechanisms of miRNA-mediated interactions and dissect their functional roles in cancer in an effort to pave the way for miRNA-based therapeutics in clinical settings.

To study the dynamics of biomolecular interaction networks, we focused on a protein-protein interaction network in living cells. Our collaborators at the School of Medicine planned to synthetically develop and characterize a biomaterial, which was produced by this protein-protein interaction network, and which would act as a molecular sieve to control the passage of biomolecules in living cells. And we wanted to computationally model the formation of this biomolecular sieve, termed a hydrogel, and characterize its properties that were relevant to the experimental work.

Specifically, we investigate a novel strategy for generating intracellular hydrogels, termed iPOLYMER for intracellular Production Of Ligand-Yielded Multivalent EnhanceRs. This particular design not only circumvents invasive approaches, such as microinjection, but also enables hydrogel formation inside living cells in a rapidly inducible manner.

To overcome difficulties in evaluating gel formation *in situ*, we developed a physical computational model for three-component multivalent-multivalent molecular interactions that led to a rigorous method for computationally implementing iPOLYMER. Our approach was based on a realistic kinetic Monte Carlo simulation algorithm that produced sufficiently accurate approximations of stochastic reaction-diffusion dynamics. In particular, we spatially discretized the well-known continuous-space Doi model of stochastic

reaction-diffusion, and obtained a physically valid approximation based on the reaction-diffusion master equation (RDME). This led to a Markov process model that describes the time evolution of the location of each basic or aggregate molecule at a resolution of one voxel in the system. We simulated the resulting process by a stochastic kinetic Monte Carlo algorithm.

In addition, we employed our computational model of iPOLYMER to assess the effects of various parameters on gel formation and its properties. This presented us and our experimental collaborators with a deeper understanding of the problem of gel synthesis, which guided the experimental design and provided further validation of the experimental findings and conclusions.

In the end, we would like to note that the findings from all three research problems we tackled here present strong computational evidence that proves to be highly valuable to experimental biologists in providing reliable predictions and systemic insights which could help them guide their applied research with an efficient and cost-effective approach.

Dissertation Committee:

Primary Reader: Dr. John Goutsias, Department of Electrical and Computer Engineering

Secondary Reader: Dr. Daniel Naiman, Department of Applied Mathematics and Statistics

Dr. Howard Weinert, Department of Electrical and Computer Engineering

Dr. Trac Tran, Department of Electrical and Computer Engineering

Dr. Takanari Inoue, Institute of Basic Biomedical Sciences, School of Medicine

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. John Goutsias, for his kind support, genuine mentorship, and constant encouragement during the course of my PhD studies. He has been a superb mentor and a really great advisor, who patiently taught me to become a better scientist, researcher and writer. He generously allowed me to follow my passion during my time at Hopkins. Without John, this dissertation and work simply would not have been possible. I consider myself incredibly fortunate for having the opportunity to work with him as my advisor and mentor in this period. Thank you, John!

I also want to express my appreciation to the other members of my dissertation committee, Dr. Daniel Naiman, Dr. Howard Weinert, Dr. Trac Tran and Dr. Takanari Inoue for their time, discussions, suggestions and comments. I am honored they accepted to be on my dissertation committee. Thank you, Dr. Naiman, for your kind support and allocating the time to discuss the work together and accepting to be the Second Reader of my dissertation. This really is an honor to me.

I would like to thank Dr. Takanari Inoue at the School of Medicine for giving me

this unique opportunity to collaborate with him and his group on this incredibly exciting project for the past three years. We have had so many joyful scientific discussions on our collaboration and I want to thank him for all his support and always being enthusiastic about our work. I could not have asked for a better collaborator. Your friendship, good humor, and knowledge was very motivating to me and was fundamental to this work. Thank you, Dr. Inoue! I also would like to thank Shiva, Hideki and other members from Dr. Inoue's lab. Shiva initially made the introduction on this collaborative work, and Hideki provided very valuable feedback on the work and we had several insightful discussions together.

I am also grateful to Dr. Mollie Meffert for giving me the opportunity to work with her and her group members. The collaboration with Mollie's group was on a high impact translational research work and I am very fortunate to have had the opportunity to contribute towards this research as an application of my computational work on microRNA networks. Working with an experienced experimental group closely was highly valuable to me and I learned a lot from this collaborative work. I also would like to thank Josh, Megha, Christina and other members from Mollie's lab for the discussions and feedback which added a lot of value to this work.

Special thanks to Dr. Rachel Green who is a pioneer in this area of research, for her deep understanding of our work and all the discussions, support and helping me connect with an experimental group to pursue a collaborative research work. Rachel's evaluation of our research meant a lot to me and has been exceptionally encouraging to me given her brilliant record of research as one of the leaders in this field in the world.

ACKNOWLEDGMENTS

I also would like to thank Drs. Luigi Marchionni, Leslie Cope and Elana Fertig for teaching me the skills that were essential in conducting this research. Their help and support was very important and made this work much stronger.

I want to thank Dr. Youseph Yazdi for his support and mentorship on all professional and scientific subjects. His advice and suggestions have been extremely helpful and enlightening on many occasions.

Special thanks to Garret Jenkinson, my lab mate, who helped me a lot with many things given he was ahead of me in the program and very kindly shared his experience with me during the course of my PhD, which was extremely helpful. Garret, you are a fantastic researcher and teacher, and I am sure you will have a bright academic future with all your significant achievements.

To Valentina and other past and present members of CIS family, thank you for your support and friendship. You are a remarkable group of colleagues, with whom I have had many fruitful discussions about scientific and non-scientific topics.

I would also like to thank Dr. Casey Overby for her kind support during the last couple of months in my PhD studies. This is highly valuable to me.

My most heartfelt appreciation goes to my dear parents, and wonderful brother, who always supported me, loved me, and encouraged me in my academic pursuits. Undoubtedly, without their unconditional love, I would not be where I am today. Thank you for your patience and taking this journey with me. You all mean a lot to me!

Dedication

This dissertation is dedicated to my mother, Beti, my dad, Hassan, and my brother, Amir.

Contents

List of Tables

List of Figures

This page is intentionally left blank.

List of Genes and Proteins

Gene Name	Description
AKT	Gene encoding RAC-alpha serine/threonine-protein kinase
APP	Gene encoding amyloid beta precursor protein
AR	Androgen receptor
BBC3	BCL2 binding component 3
BDNF	Brain derived neurotrophic factor
CDH1	Cadherin 1
CDK2	Cyclin dependent kinase 2
CDKN1B	Cyclin dependent kinase inhibitor 1B
DICER	Dicer 1, ribonuclease III
E12/E47	Gene encoding transcription factor 3 (TCF3)
E2F1	E2F transcription factor 1
E2F2	E2F transcription factor 2
ELK1	Gene encoding a member of ETS transcription factors (ELK1)
ERG	Gene encoding a member of ETS transcription factors (ERG)
ERK	Extracellular regulated MAP kinase
ETV4	ETS variant 4
FMR1	Fragile X mental retardation 1
GATA3	GATA binding protein 3
IGF1R	Insulin like growth factor 1 receptor
LIN28a	Lin-28 homolog A
MYC	V-myc avian myelocytomatosis viral oncogene homolog
NKX3-1	NK3 homeobox 1
NOXA	A pro-apoptotic member of the Bcl-2 protein family
PI3K	Phosphatidylinositide 3-kinases
PIK3R1	Phosphoinositide-3-kinase regulatory subunit 1

LIST OF GENE/PROTEIN NAMES AND THEIR DESCRIPTIONS

Gene Name	Description
PTEN	Phosphatase and tensin homolog
RAF	Gene encoding a proto-oncogene serine/threonine-protein kinase
RAS	An oncogene encoding a family of proteins
RB1	Gene encoding retinoblastoma protein 1
SMAD1	Gene encoding SMAD 1 from SMAD family of proteins
SMAD3	Gene encoding SMAD 3 from SMAD family of proteins
SNAI1	Snail family transcriptional repressor 1
SNAI2	Snail family transcriptional repressor 2
SRF	Serum Response Factor
TARP	Gene encoding TCR gamma alternate reading frame protein
TGFB	Gene encoding transforming growth factor beta
TGFB1	Gene encoding transforming growth factor beta 1
TGFB2	Gene encoding transforming growth factor beta 2
TGFB3	Gene encoding transforming growth factor beta 3
TGFBR1	Gene encoding transforming growth factor beta receptor 1
TGFBR2	Gene encoding transforming growth factor beta receptor 2
TGFBR3	Gene encoding transforming growth factor beta receptor 3
TMPRSS2	Gene encoding transmembrane protease, serine 2
TP53	Tumor protein p53
TRBP	RISC loading complex RNA binding subunit (TARBP2)
TRKB	Neurotrophic receptor tyrosine kinase 2
TWIST	Twist family bHLH transcription factor 1
WIF1	WNT inhibitory factor 1
WNT	Proto-oncogene protein Wnt-1
ZEB1	Zinc finger E-box binding homeobox 1
ZEB2	Zinc finger E-box binding homeobox 2

LIST OF GENE/PROTEIN NAMES AND THEIR DESCRIPTIONS

Protein Name	**Description**
E2F	The family of E2F transcription factors
FKBP	FK506-binding protein 5
FMRP	Fragile X mental retardation protein
FRB	FKBP12-rapamycin binding protein
LIN28	A family of RNA-binding proteins
TGF-β	Transforming growth factor beta family of proteins
TGN38	Trans-Golgi network integral membrane protein

Note: In this Dissertation, symbols for genes and RNAs are italicized whereas symbols for proteins are not. This formatting convention is applied to the genes and proteins referred to in the text. In the Figures, it will be evident whether a name refers to a gene or protein based on the particular context.

Chapter 1

Introduction

For more than the entire past century, classical experimental methodologies have dominated biological research, providing a wealth of information about individual molecular species in cells and their functions. However, there is an increasing and strong level of evidence suggesting that an isolated biological function can only rarely be attributed to an individual biological molecule. Instead, more recently, it is argued that most biological characteristics are due to complex interactions between the cell's numerous constituents, such as proteins, DNA and RNA. Therefore, a major challenge for the biological sciences in this century is to unravel the structure and the dynamics of these complex intracellular interactions at a systems level. Systems biology, being an interdisciplinary biology-based area of research, concentrates on such systemic understanding of the complex interaction networks in biological systems by utilizing the computational and statistical methodologies applied to biological and medical data.

The behavior of most complex systems, including the cellular interaction networks, originates from the orchestrated interplay of many components that interact with each other through pairwise interactions. And, in order to understand the function of a cell, it is, in this way, convenient to conceptualize the cellular activities as systems of interacting elements. For such a systems-level representation, one needs to know: i) the identity of the components that constitute the biological system of interest, ii) the interactions among these components, and iii) the dynamic behavior of these molecular entities, meaning how their abundance or activity changes over time in various conditions [3, 75]. It is noteworthy that early attempts at systems-level understanding of biology suffered from inadequate data to be utilized to base the relevant mathematical theories and computational models upon; however, the emergence of high-throughput technologies and high-precision experimental techniques brought an abundance of data on system elements and interactions, leading to a revival of the field of systems biology. In particular, these experimental methods enable the measurement of expression levels for thousands of genes and the determination of thousands of protein-protein interactions in cells.

It is well known that the creation of models of the functions of genes and proteins in cells is of fundamental and immediate significance to the emerging field of computational systems biology. Some of the most successful attempts at cell-scale modeling to date have been based on constructing networks that represent hundreds of experimentally-validated biochemical interactions, while others have been very successful at inferring statistical networks from large amounts of high-throughput data. These types of cellular networks

(metabolic, regulatory, or signaling) can be analyzed, and predictions about biological behavior made and tested. Many types of statistical and computational models have been built and applied to study cellular behavior and in this research work, we focus on two distinct instances, one from each of the two broad types of models used in computational systems biology: i) statistical inference models applied to gene regulatory interaction networks and ii) biochemical reaction models applied to protein-protein interaction networks. It is helpful to note that, in both instances of the models we consider here, the interaction of system elements is modeled by a graph representation which ascribes the graph nodes (also called vertices) to the molecular species of interest and represents their pairwise relationships by edges (also called links) connecting pairs of nodes. The nodes of (sub)cellular systems may be genes, mRNAs, proteins, or other molecules. Directed edges have a specified source (starting) node and target (end) node and are most suited to represent regulatory relationships. Non-directed edges are most appropriate for mutual interactions, such as protein-protein binding or for relationships whose source and target are not yet determined.

It goes without saying that computational models and methods in systems biology are most useful if they lead to concrete and novel biological predictions that could guide the experimental biologists in validating the resulting predictions and if they address the pressing issues in biology and medicine, with the potential to have a high impact in applied research. As a matter of fact, the three specific research problems we tackle here (which are categorized under the previous two broad modeling paradigms), have been identified as critical, applied research problems by our collaborators at the Johns Hopkins School of Medicine.

The first two problems, i.e. integrative identification of microRNA-mediated gene regulatory interaction networks in the contexts of: i) prostate cancer, and ii) autism spectrum disorders, are tackled by the statistical inference models, and the third research problem, i.e. modeling the dynamics of synthetic protein-protein interaction networks in living cells, is tackled by the biochemical reaction system modeling techniques. In the following, we briefly review the biological and computational backgrounds related to these three specific problems and their significance in biology and medicine. In the first two research problems we tackle in this work, our focus will be on identifying microRNA-mediated networks in gene regulatory interactions networks.

1.1 MicroRNA/TF-mediated Networks in Prostate Cancer

MicroRNAs (miRNAs) have attracted a great deal of attention in biology and medicine. They are small non-coding ribonucleic acids (RNAs) that extensively regulate gene expression in metazoan animals, plants and protozoa. Approximately 22 nucleotides in length, miRNAs usually repress gene expression by binding to sequences with partial complementarity on target messenger RNA (mRNA) transcripts. In mammals, miRNAs are thought to control the activity of more than 60% of all protein-coding genes and extensively participate in the regulation of many cellular functions [37, 114]. It has been hypothesized that miRNAs interact with transcription factors (TFs) in a coordinated fashion to play key roles

in regulating signaling and transcriptional pathways and in achieving robust gene regulation.

In our first research work, we propose a novel integrative computational method to infer certain types of deregulated miRNA-mediated regulatory circuits (motifs) at the transcriptional, post-transcriptional and signaling levels. The first set of motifs that our method considers are three-node feed-forward loops (FFLs) that have recently attracted a great deal of attention among systems and experimental biologists. These motifs are excellent models of coordinated miRNA-mediated and transcriptional regulation, which have been hypothesized to be prevalent in the human and mouse genomes [155].

In addition to the modulatory and/or reinforcing gene regulatory roles that miRNAs are known to play in concert with TFs in the context of FFLs, they have been hypothesized to play key roles in regulating signaling pathways as well. In this respect, although miRNAs are known to have subtle effects on protein levels of individual targets, their cumulative influence can significantly affect the outcomes controlled by signaling pathways, given the multiplicity of their targets and concurrent downregulation of several of these targets. To take this important aspect into account, our method also considers the basic Type III loop motif, in which a miRNA targets two gene transcripts, G-1 and G-2, whose proteins could potentially interact with each other according to a pathway map provided in the KEGG database (http://www.kegg.jp). See Fig. 2.1.

To reliably predict miRNA-target interactions from mRNA/miRNA expression data, our method collectively utilizes the aforementioned molecular structures identified to be

statistically over-represented in gene regulatory networks (FFLs and Type III loops), sequence-based miRNA-target predictions obtained from several algorithms, known information about mRNA and miRNA targets of TFs available in existing databases, available molecular sub-typing information, and state-of-the-art statistical techniques to appropriately constrain the underlying analysis. In this way, the method exploits almost every aspect of extractable information in the expression data.

We apply our procedure on mRNA/miRNA expression data from prostate tumor and normal samples and detect numerous known and novel miRNA-mediated deregulated loops and networks in prostate cancer. We also demonstrate instances of the results in a number of distinct biological settings, which are known to play crucial roles in prostate and other types of cancer. Our findings show that the proposed computational method can be used to effectively achieve notable insights into the poorly understood molecular mechanisms of miRNA-mediated interactions and dissect their functional roles in cancer in an effort to pave the way for miRNA-based therapeutics in clinical settings.

1.2 MicroRNA/TF-mediated Networks in Autism Spectrum Disorders

As our second research problem, we look into the role of LIN28-regulated miRNAs and their interaction networks in Autism Spectrum Disorders (ASDs). In doing so, we are going to utilize the miRNA-mediated Type III loops and also FFLs we introduced in our

first research work. The reason miRNAs could potentially play a crucial role in ASDs is because, it is known that the regulation of gene expression at the level of translation is a critical factor in the neuronal response to several stimuli, including synaptic activity [60] and neurotrophins [130], among others. Although many such stimuli enhance the overall synthesis of cellular proteins, their responses demonstrate selection of specific mRNAs for enhanced translation. One of the well-studied examples of such stimuli is the brain-derived neurotrophic factor (*BDNF*), which is broadly expressed in the mammalian brain, and critically contributes to modifications of synaptic growth and function.

The effects of BDNF on protein synthesis selectively targets a minority of expressed mRNAs (an estimated 4% or less of expressed mRNAs which undergo enhanced translation [130]). In [59], it was experimentally determined that the function of miRNA biogenesis pathways plays a pivotal role in BDNF-mediated regulation of translation.

In particular, it is reported that BDNF induces widespread changes in miRNA biogenesis by rapidly elevating the miRNA processing enzype, DICER, which increases mature miRNA levels. Moreover, BDNF induces LIN28a, which is a protein that prevents the processing of certain miRNAs, and as a result, causes the loss of the corresponding mature miRNAs and a consequent upregulation in translation of their target mRNAs. In this way, it was shown that target specificity of BDNF-induced translation is achieved by a two-part mechanism, i.e. the combined action of BDNF on DICER and LIN28a protein [59]. This dual mechanism, depicted in Fig. 1.1, results in genome-wide control of translation specificity that involves BDNF-dependent positive and negative regulation of miRNA biogenesis

pathway.

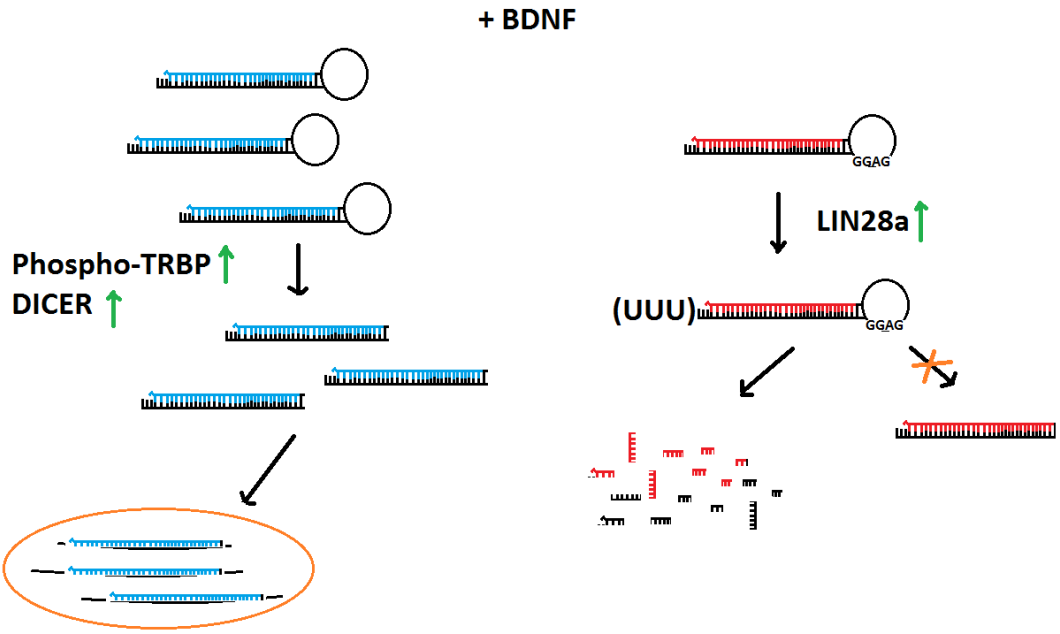

Figure 1.1: The dual mechanism of BDNF inducing both positive (left path) and negative (right path) regulation of miRNA biogenesis through DICER and LIN28a proteins. In one mechanism (on the left), BDNF induces the phosphorylation of TRBP, which leads to the elevation of DICER levels. Elevated DICER levels could invoke mature miRNA biogenesis, and with elevated levels of miRNAs, additional mRNAs could be targeted for repression (shown at the bottom of the left path). In the second mechanism, with BDNF exposure, it is experimentally observed that a rapid and transcription-independent increase in LIN28a levels takes place. In addition, LIN28a recognizes a functionally confirmed "GGAG" sequence motif on the terminal loop of certain pre-miRNAs. This causes the uridylation of the molecule, indicated by (UUU) on the right path, and as a result, it suppresses processing of the targeted pre-miRNA to the mature miRNA. Adopted from [59].

Although it is possible that alternative mechanisms could coexist with the one discussed above, the experimental results obtained in [59] strongly suggest that dual control by BDNF of the miRNA biogenesis pathway by means of LIN28a and DICER plays a crucial role in selectively determining both upregulated and downregulated targets in BDNF-mediated

translation.

In this combined computational-experimental research work, our goal was to validate the biological hypothesis that pathological regulation of LIN28-regulated miRNAs and their network of interactions may lead to a selective overabundance of growth-promoting synaptic proteins. We constructed these networks of interactions by identifying the Type III loops and FFLs we introduced in our first research work. The results we obtained in this way could account for synaptic and cognitive functions in FXS, and could help us infer the miRNA-mediated systemic insights governing this dysregulation.

1.3 Modeling Synthetic Protein-Protein Interaction Networks in Living Cells

The former two research problems were examples of statistical network inference in systems biology. In our third research work, our goal was to model the dynamics of a protein-protein interaction network in living cells, as an instance of the second broad category of research in systems biology. For this purpose, we developed a combined computational method by utilizing certain biochemical reaction modeling techniques. Our collaborators at the School of Medicine planned to synthetically develop and characterize a biomaterial, which was produced by this protein-protein interaction network, and which would act as a molecular sieve to control the passage of biomolecules in living cells. And we wanted to computationally model the formation of this biomolecular sieve, termed a

hydrogel, and characterize its properties that were relevant to the experimental work.

Specifically, we investigate a novel strategy for generating intracellular hydrogels, termed iPOLYMER for intracellular Production Of Ligand-Yielded Multivalent EnhanceRs. This particular design not only circumvents invasive approaches, such as microinjection, but also enables hydrogel formation inside living cells in a rapidly inducible manner.

To overcome difficulties in evaluating gel formation *in situ*, we first developed a computational model of iPOLYMER to assess the effects of various parameters on gel formation and its properties. This presented us and our experimental collaborators with a deeper understanding of the problem of gel synthesis, which guided the experimental design and provided further validation of the experimental findings and conclusions. Our collaborators succeeded in observing punctate polymer aggregates rapidly induced in living cell by practicing iPOLYMER, and examined their biophysical characteristics including the turnover rate and molecular sieving effects. They additionally reconstituted the gel formation *in vitro* using purified proteins, and described its characteristics as a molecular sieve. Isolation of the gel *in vitro* also confirmed the identity of the material as a hydrogel that retains water inside. Moreover, the strong potential of iPOLYMER was demonstrated by synthesizing biologically functional entities, such as a size-dependent diffusion barrier and a nucleation platform for RNA molecules.

1.4 Contributions of the Dissertation

Here, we would like to note that the three research problems we address in this dissertation, involve two main instances of the two broad areas of research in systems biology (one from each), i.e., i) statistical inference of biological interaction networks, and ii) the dynamic modeling of such networks. In our first research work, we propose IntegraMiR, an integrative computational method to reliably predict miRNA-target interactions from mRNA/miRNA expression data, utilize certain molecular structures identified to be statistically over-represented in gene regulatory networks (FFLs and Type III loops), sequence-based miRNA-target predictions obtained from a set of reliable algorithms, known information about mRNA and miRNA targets of TFs available in existing databases, available molecular subtyping information, and state-of-the-art statistical techniques to appropriately constrain the underlying analysis.

By comparison, the methods proposed in this specific research area by other groups, that are reviewed in [17, 174], each suffer from a combination of these issues: i) They are constrained to a specific motif, ii) They do not discriminate between coherent and incoherent FFLs, which is required for a systems-level understanding of transcriptome changes in disease, iii) The standard statistical tests used to identify differentially expressed genes between two conditions in a typical gene expression profiling study, as adopted by previous methods [17, 174], become fundamentally flawed in the presence of unaccounted sources of variability (due to biological and experimental factors among others) [15, 86, 87]. Molecular subtyping information is a critical example of such sources of variability, iv) The pre-

diction problem is defined on all edges of a given motif, and this substantially lowers the accuracy of the predicted miRNA-target interactions, and will make the predictions highly unreliable to be utilized by experimental biologists. Our proposed integrative method, IntegraMiR, addresses all the above issues and is specifically designed to achieve reliable predicted miRNA-target interactions with a systems perspective, that can readily be used by experimental biologists to guide their experimental procedures in a systematic way.

In our second research work, as an application of our previous work on miRNA/TF loop and network identification, we utilize the IntegraMiR paradigm and exploit additional publicly available databases and datasets to infer miRNA-mediated regulatory networks in Autism Spectrum Disorders, that were of interest to our experimental collaborators at the School of Medicine. In particular, we use certain databases and datasets that list the genes and proteins that are relevant in the context of Fragile X Syndrome, a disease in the category of ASDs, and also provide the expression levels of these genes and proteins in the disease state. In addition, our experimental collaborators provided a set of experimental data they obtained in their lab which could serve to validate part of our computational predictions and results, and help them make informed decisions on which set of genes and proteins they could focus as the most likely players in the disease state.

In this way, for this second research work, we were able to construct predicted miRNA-mediated networks using the IntegraMiR paradigm, and perform an additional level of computational analysis that allowed us to validate the relevance of certain miRNAs in the disease. Specifically, we performed gene set enrichment analysis on the predicted targets

of miRNAs of interest and validated at the computational level (in addition to the subsequent supporting experimental results) that the miRNAs our collaborators focused on in the experimental work could indeed contribute to the disease state. Combining an integrative framework similar to IntegraMiR with this additional gene set enrichment analysis to validate the relevance of the miRNAs of interest is a unique procedure and is essentially utilizing all the major sources of information available to us to take into account a given experimental setting to tailor the results for that specific setting.

In our third research problem, we developed a physical computational model for three-component multivalent-multivalent molecular interactions that led to a rigorous method for computationally implementing iPOLYMER. Our approach was based on a realistic kinetic Monte Carlo simulation algorithm that produced sufficiently accurate approximations of stochastic reaction-diffusion dynamics. A directly related method has recently been proposed in [89]. Although, at a first glance, this method seems to be similar to ours, there are some major and important differences. The method proposed in [89] and other recent methods discussed in Chapter 4 each suffer from a combination of the following major issues: i) The method cannot be directly related to any physical model (such as the Doi model) for multivalent binding/unbinding interactions in continuous time/space. ii) It is not clear whether the method converges to a continuous time/space model as the time-step size and the voxel volume decrease towards zero. As a consequence, the method in [89] leads to an *ad hoc* algorithm that cannot be directly related to a physical model for multivalent molecular binding. This deficiency can seriously compromise the utility and accuracy of

this method in an experimental setting. iii) They cannot provide a graphical representation of molecular aggregates. As a consequence, such methods cannot be used to study sieving properties of molecular aggregates, e.g., by means of computing Pore Size Distributions (PSDs).

Our proposed computational method models three-component multivalent-multivalent molecular interactions and provides a rigorous discretization of the well-known continuous time/space Doi model for systems that involve reactions among different types of molecules as well as molecular diffusions. In an effort to guarantee that the resulting discretization converges to the Doi model, as the voxel volume approaches zero, our method provides appropriate formulas for the probability rates of the underlying binding/unbinding reactions as well as for the probability rates of molecular diffusion. Moreover, the proposed method treats time as a continuous variable and results in a realistic kinetic Monte Carlo simulation algorithm that is expected to produce sufficiently accurate approximations of stochastic reaction-diffusion dynamics. In addition, our model considers reactions among the binding sites of the underlying reactant molecules and can thereby construct a graph representing each molecule of interest, and allow for further investigation of the sieving properties of molecular aggregates.

In the end, we would like to note that the findings from all three research problems we tackled here present strong computational evidence that proves to be highly valuable to experimental biologists in providing reliable predictions and meaningful insights which could help them guide their applied research in a systematic, efficient and cost-effective

manner.

Chapter 2

MiRNA/TF-mediated Networks in Prostate Cancer

Introduction

MicroRNAs (miRNAs) are small non-coding ribonucleic acids (RNAs) that extensively regulate gene expression in metazoan animals, plants and protozoa. Approximately 22 nucleotides in length, miRNAs usually repress gene expression by binding to sequences with partial complementarity on target messenger RNA (mRNA) transcripts. In mammals, miRNAs are thought to control the activity of more than 60% of all protein-coding genes and extensively participate in the regulation of many cellular functions [37, 114].

With few exceptions, metazoan miRNAs base-pair with their targets imperfectly, following a set of rules that have been formulated by employing experimental and bioinformatics-

based analyses [8]. This limited complementarity makes the task of computationally identifying miRNA targets very challenging and usually leads to large numbers of, mostly false, potential targets.

Earlier computational tools have mainly focused on dissecting individual miRNA-target interactions by relying on sequence-based identification of miRNA-target binding sites or on mRNA/miRNA expression data analysis [51, 126, 160]. Alternative methods use miRNA host genes as proxies for measuring the expression of embedded miRNAs [44] or employ an information-theoretic approach to identify candidate mRNAs that modulate miRNA activity by affecting the relationship between a miRNA and its target(s) [143]. On the other hand, recent work considers co-expression analysis, by assuming that targets of a given miRNA are co-expressed, at least in certain tissues or conditions [43].

Conventionally, many computational methods developed for miRNA-target prediction rely on the assumption that there is an inverse correlation between the expression level of a miRNA and that of its target [140]. However, it has been recently shown that both positive and negative transcriptional co-regulation of a miRNA and its targets are prevalent in the human and mouse genomes [133, 155]. In particular, two types of regulatory circuits (that we will be discussing shortly) have been proposed for miRNA-mediated interactions, which ascribe modulatory and/or reinforcing roles to miRNAs in their networks based on motifs, such as feed-forward loops (FFLs) [3]. As a consequence, miRNA-target predictions solely relying on an inverse correlation assumption are expected to be limited if the prediction method does not appropriately incorporate the underlying FFL network structure.

Based on the previous paradigm, several researchers have investigated the statistical over-representation of network structures involving miRNA and TF co-regulation of mRNAs to identify enriched network motifs and/or assess their prevalence in different biological contexts [19, 21, 88, 119, 141, 145, 154, 178]. Essentially, these methods compute measures of coordinated gene co-regulation by miRNA and TF regulators. Other investigators have considered regression methods or Bayesian models to quantify statistical associations by determining changes in the expression level of a given mRNA explained by the expression levels of TFs and miRNAs predicted to target the mRNA based on sequence information [83, 132, 176, 177]. Subsequently, they use the inferred relationships to delineate significant network structures and motifs in a fashion similar to that employed in the aforementioned methods. It is important to note however that the collective findings produced by all these approaches provide further support for the importance of miRNA/TF-mediated FFLs as prevailing network motifs across different biological contexts, reconfirming the hypotheses originally proposed in [133, 155].

In addition to the above, disruptions in gene regulation (for instance, by genetic and epigenetic alterations) believed to induce changes in normal cell function that lead to the progression of pathological conditions, such as cancer, are disseminated through gene regulatory networks. As a consequence, effective treatment of many human diseases may require a fundamental and systemic understanding of genomic regulators, such as miRNAs and TFs, and their networks of interaction. However, systematically inferring molecular interactions by experimental methods is both difficult and costly. Therefore, it is highly de-

sired to develop "reliable" computational approaches capable of identifying such networks. Network predictions can subsequently be used by an expert biologist to formulate novel hypotheses and effectively proceed with their experimental investigation and validation.

Recently, several new methods have been proposed for identifying coordinated miRNA/TF interactions [17, 174]. However, for a given motif structure (e.g., an FFL), these methods attempt to predict the underlying interactions (the three edges of an FFL) by utilizing limited biological information and a narrow set of computational tools. As a result, although the methods are effective in providing insights into the prevalence of various motif instances in gene regulatory networks, they may not produce reliable predictions from an experimental perspective.

The performance of some of the previous methods has been recently tested in [17]. It was observed that, although some methods were capable of achieving a reasonable success rate in predicting instances of one type of interaction, they were less effective in predicting instances of the other two types, with several algorithms having a success rate of close to or less than 1% in predicting TF-mRNA and TF-miRNA interactions. This highlights the critical fact that predicting pair-wise molecular interactions and constructing higher-order instances of motifs using the predicted edges could translate to higher overall false-positive rates. Since there is a wealth of information on how a TF binds its targets and on their specific regulatory roles, we decided to consider only *experimentally* validated TF-mRNA and TF-miRNA interactions under the FFL framework and shift focus on reliably predicting the poorly understood miRNA-target interaction edge. We believe that, by appropriately

constraining the underlying statistical analysis problem, we could potentially increase the reliability of miRNA/TF-mediated gene regulatory loop predictions.

To further constrain the miRNA-target interaction prediction problem, we focus, in this work, on certain three-node regulatory motifs. The first set of motifs that our method considers are three-node FFLs that have recently attracted a great deal of attention among systems and experimental biologists. These motifs are excellent models of coordinated miRNA-mediated and transcriptional regulation, which have been hypothesized to be prevalent in the human and mouse genomes [155].

We consider two Type I FFL motifs, in which the miRNA and TF are the upstream and downstream regulators, respectively, as well as four Type II FFL motifs, in which the TF is now the upstream regulator, whereas the miRNA is the downstream regulator – see Fig. 2.1. From a mechanistic perspective, these six FFLs are classified as being *coherent* or *incoherent*. In the coherent case, the miRNA and TF regulators act in a coordinated fashion to reinforce the regulation logic along two feed-forward paths. In Type I and Type II-B coherent FFLs, these paths simultaneously repress the expression of the targeted mRNA. The resulting mechanism is used, for instance, to subdue leaky transcription of a gene by ensuring that its expression stays at an inconsequential level. On the other hand, in a Type II-A coherent FFL, the TF reinforces the transcription of the targeted mRNA by directly activating it as well as by inhibiting its repression by the targeting miRNA regulator.

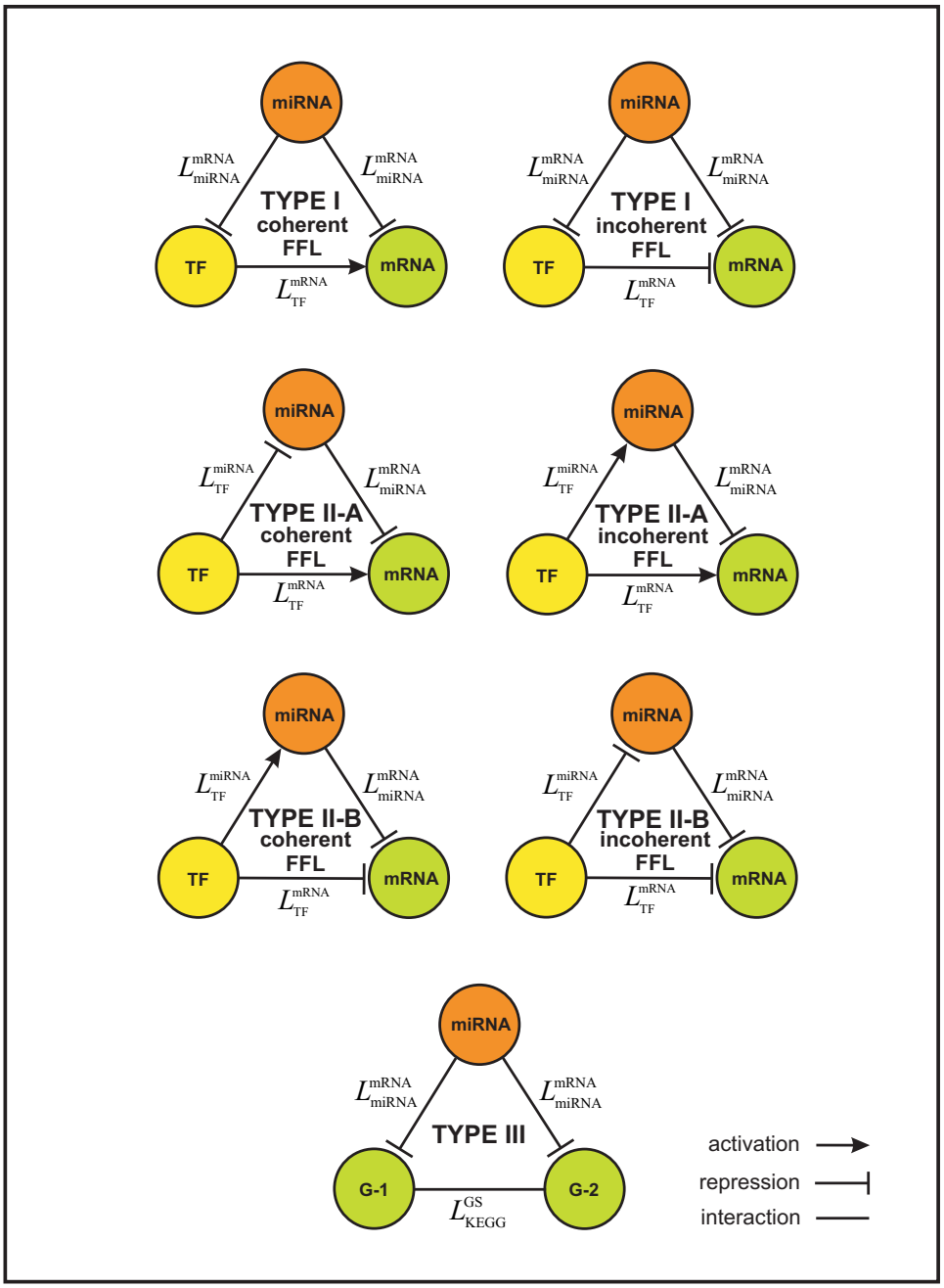

Figure 2.1: Three-node regulatory motifs considered by IntegraMiR. The Type I FFL consists of triplets (miRNA, TF, mRNA) such that a miRNA simultaneously targets a mRNA and its TF mRNA. The Type II FFL consists of triplets (miRNA, TF, mRNA) such that a TF simultaneously regulates a miRNA and its target mRNA. Finally, the Type III loop consists of triplets (miRNA, G-1, G-2) such that the miRNA simultaneously targets two transcripts in a given KEGG pathway, one from each gene G-1 and G-2, whose corresponding proteins could potentially interact with each other based on a pathway map provided in the KEGG database. The labels on the edges of these motifs are defined in Table 2.1.

In the incoherent FFLs, the miRNA and TF regulators act in a coordinated fashion to fine-tune the expression of the targeted mRNA. More specifically, any deviation from the steady-state concentration of the upstream regulator (i.e., the miRNA in Type I and the TF in Type II-A and Type II-B FFLs) would drive the targeted mRNA, as well as the downstream regulator, away from their steady-state levels in the same direction. In this way, the downstream regulator can balance the expression of the targeted mRNA, compensating fluctuations in the expression level of the upstream factor.

Certain cellular processes might be ultra-sensitive to the activity of a given transcript in a specific biological context. In these situations, the "noise buffering" mechanism provided by incoherent FFLs helps maintain target protein homeostasis and ensures that an uncoordinated drift from the steady-state level of the upstream regulator may not result in an undesirable variation in the target protein level which can lead to pathological outcomes. MiRNAs are particularly effective in this setting, owing to their rapid mechanism of action at the post-transcriptional level, as opposed to transcriptional repressors, thus accelerating noise buffering [155].

In addition to the coherent and incoherent FFLs, our method also takes into account the basic Type III loop motif depicted in Fig. 2.1, in which a miRNA targets two gene transcripts, G-1 and G-2, whose proteins could potentially interact with each other according to a pathway map provided in the KEGG database (http://www.kegg.jp). The existence of Type III loop motifs is supported by two key hypotheses: (i) miRNAs play major roles in regulating signaling pathways due to their sharp dose-sensitive nature [9, 26, 62, 128, 173],

and (ii) targets of single miRNAs are more connected (i.e., interact) at the protein level than expected by chance [26, 57, 92, 156].

By comparison, the method proposed in [174] considers only Type II FFLs and does not discriminate between coherent and incoherent FFLs, which is required for a systems-level understanding of transcriptome changes in disease. Moreover, the standard statistical tests used to identify differentially expressed genes between two conditions in a typical gene expression profiling study, as adopted by previous methods [17, 174], become fundamentally flawed in the presence of unaccounted sources of variability (due to biological and experimental factors among others) [15, 86, 87]. Molecular subtyping information is a critical example of such sources of variability.

To reliably predict miRNA-target interactions from mRNA/miRNA expression data, we propose an integrative method to collectively utilize the aforementioned molecular structures identified to be statistically over-represented in gene regulatory networks (FFLs and Type III loops), sequence-based miRNA-target predictions obtained from several algorithms, known information about mRNA and miRNA targets of TFs available in existing databases, available molecular subtyping information, and state-of-the-art statistical techniques to appropriately constrain the underlying analysis. We discuss the proposed method in detail in the next section.

2.1 An Integrative MiRNA-mediated Network Analysis Method

To address the previous issues in inferring miRNA-target interaction networks, in this research work, we develop IntegraMiR, a novel integrative analysis method that can be used to infer certain types of regulatory loops of deregulated miRNA/TF interactions which appear at the transcriptional, post-transcriptional and signaling levels in a statistically over-represented manner.

In the context of our first research problem, the proposed method assigns biological roles to miRNAs by integrating five major sources of information together with state-of-the-art statistical techniques to reliably infer specific types of miRNA-target interactions in the context of regulatory loops. In particular, IntegraMiR utilizes:

(i) mRNA and miRNA expression data.

(ii) Sequence-based miRNA-target information obtained from different algorithms.

(iii) Known information about mRNA and miRNA targets of TFs available in existing databases.

(iv) Certain three-node motifs in gene regulatory networks.

(v) Known molecular subtyping information available with gene expression data.

To do so, IntegraMiR identifies deregulated miRNAs, TFs and mRNAs by performing statistical analysis within a constrained framework that uses “prior” information comprising

recently discovered motifs, available knowledge on miRNA/mRNA transcriptional regulation, and known protein-level interactions on signaling pathways. To illustrate the effectiveness and potential of this method, we apply it on mRNA/miRNA expression data from tumor and normal samples and identify several known and novel deregulated loops in prostate cancer (PCa). This allows us to demonstrate instances of the results and findings in a number of distinct biological settings, which are known to play crucial roles in PCa and other types of cancer.

The flow-chart depicted in Fig. 2.2 provides a general description of the different steps employed by IntegraMiR.The procedure uses mRNA and miRNA expression data obtained from prostate tissue at two different biological conditions (normal vs. cancer). It moreover employs results obtained by sequence-based miRNA target prediction algorithms and incorporates information extracted from four databases available online, namely:

- mSigDB (www.broadinstitute.org/gsea/msigdb)
- miRTarBase (http://mirtarbase.mbc.nctu.edu.tw)
- TRANSFAC[1] (www.gene-regulation.com/pub/databases.html)
- TransmiR (http://202.38.126.151/hmdd/mirna/tf).

The first step of IntegraMiR applies standard preprocessing techniques on the raw expression data (such as background correction, normalization, and data heterogeneity cor-

[1]Recently, ENCODE released information on TF binding sites based on ChIP-seq experiments for 161 TFs in 91 cell lines (http://genome.ucsc.edu/ENCODE). Unfortunately, this database does not provide the regulation type (activation or repression) of a particular TF-target interaction, information that is critical in our approach. For this reason, IntegraMiR uses TRANSFAC. However, once this information becomes available through ENCODE or any other TF-target database, it can be readily utilized by IntegraMiR.

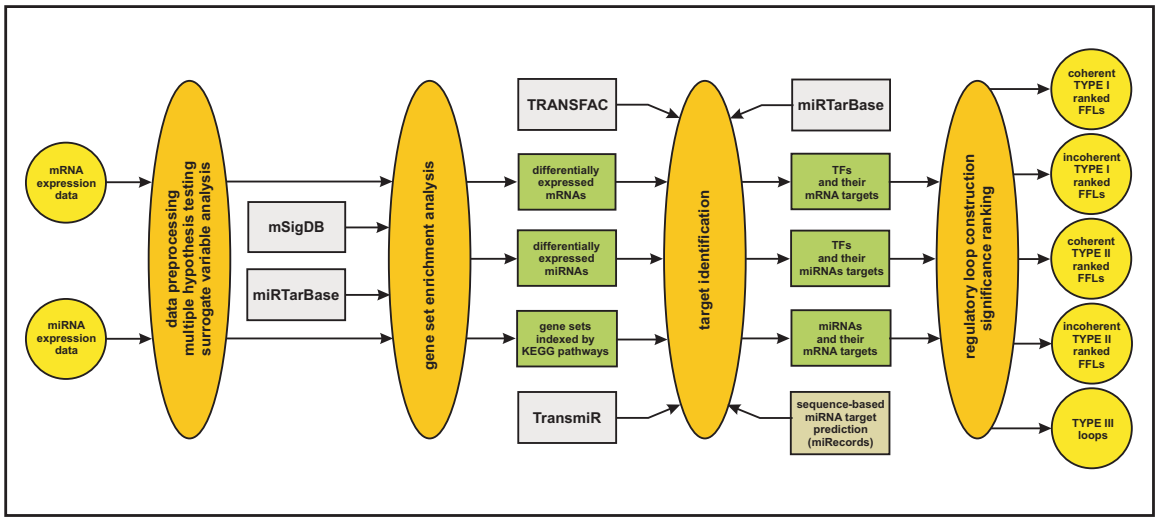

Figure 2.2: General description of IntegraMiR. The method assigns biological roles to miRNAs by integrating five major sources of information together with state-of-the-art statistical techniques to reliably infer specific types of miRNA-target interactions in the context of regulatory loops from mRNA and miRNA expression data.

rection) to improve data quality, followed by multiple hypothesis testing (MHT) and surrogate variable analysis (SVA) to identify mRNAs and miRNAs that are differentially expressed between the two biological conditions, while correcting for biological variability due to molecular subtyping, multiple testing and batch effects.

The second step implements additional statistical analysis using gene set enrichment analysis (GSEA) to further evaluate the biological significance of certain mRNAs and miRNAs that are not deemed to be differentially expressed by MHT. By employing the molecular signatures database mSigDB of annotated gene sets for use with GSEA and the *experimentally* verified miRNA target database miRTarBase, IntegraMiR constructs three separate groups of gene sets and evaluates the statistical significance of each gene set enriched for deregulation in the available mRNA expression data. The first group consists of gene sets in the mRNA data indexed by a TF mRNA that is not deemed to be differentially

expressed by MHT and is determined by mSigDB to directly regulate each gene in the gene set. The second group consists of gene sets in the mRNA data indexed by a miRNA that is not deemed to be differentially expressed by MHT and is determined by miRTarBase to target each gene in the gene set. The third group consists of gene sets in the mRNA data indexed by a specific KEGG signaling pathway [69, 70] included in mSigDB. Finally, TFs associated with statistically significant enriched gene sets are amended to the list of those mRNAs deemed to be differentially expressed by MHT to generate a combined list of differentially expressed mRNAs, and the same is done for miRNAs. We should note here that mSigDB is widely used to obtain gene sets for GSEA analysis. On the other hand, we employ MiRTarBase since this database has accumulated a relatively large number of experimentally validated miRNA-target interactions.

In brief, GSEA determines whether a given set of genes shows statistically significant concordant differences between two biological states [142]. The main reason IntegraMiR applies GSEA after the initial hypothesis testing step is to improve detection of differentially expressed TFs and miRNAs, which may be missed when single expression levels show only moderate changes between the two biological conditions. As a matter of fact, if a number of transcripts are known to participate in a common biological mechanism, then even moderate changes in the expression levels of these transcripts may be statistically significant due to the fact that known biological relationships between transcripts may result in higher statistical power when detecting small variations in their expression levels as compared to the case of single transcripts. Moreover, for certain TFs, TF mRNA

expression cannot necessarily be used as a proxy of its activity at the protein level, due to post-transcriptional and post-translational modifications of TFs [20, 85]. To address these issues, IntegraMiR also considers the collective differential expression of genes, as opposed to several procedures followed by other related work discussed earlier that mainly build their analyses on statistics obtained from single transcripts.

The third step of IntegraMiR uses the results obtained by MHT and GSEA, as well as available biological knowledge and sequence-based miRNA target predictions, to identify known *directly* regulated targets of differentially expressed TFs and miRNAs and predicted targets for the miRNAs. By employing the eukaryotic TF database TRANSFAC and the TF/miRNA regulation database TransmiR, IntegraMiR produces a list of differentially expressed TFs together with their gene targets and the regulation type (activation or repression) for each target gene. It also produces a list of differentially expressed TFs together with their differentially expressed miRNA targets and the regulation type for each target miRNA. Note that our choice for using TRANSFAC and TransmiR is based on the fact that TRANSFAC reliably provides the crucial information of regulation type (activation/repression) of a transcription factor and its target gene(s), whereas TransmiR provides the crucial information of the miRNA(s) being regulated by it. On the other hand, to identify mRNA targets of differentially expressed miRNAs, IntegraMiR employs miRecords (http://mirecords.umn.edu/miRecords), an integrated sequence-based miRNA target prediction tool, as well as miRTarBase, a database of experimentally validated miRNA targets. At this step, IntegraMiR produces a list of differentially expressed miRNAs with the

corresponding sequence-based target predictions, amended with experimentally validated mRNA targets from miRTarBase to help identify true-positive and false-negative predictions by using available biological knowledge. In this respect, IntegraMiR incorporates a *predictive* module (exploiting miRecords) and a *non-predictive module* (miRTarBase) to accomplish this task.

The fourth step of IntegraMiR implements a technique, described in Section 2.2.7, to construct deregulated loops of the types depicted in Fig. 2.1 using the results obtained from the previous steps. IntegraMiR constructs the following three types of regulatory loops:

(i) An FFL comprising a miRNA which simultaneously targets a TF and a mRNA that is directly regulated by the TF.

(ii) An FFL comprising a TF which directly regulates a miRNA and a mRNA that is directly targeted by the miRNA.

(iii) A regulatory loop comprising a miRNA which simultaneously targets two different genes in a given KEGG pathway whose proteins could potentially interact with each other based on a pathway map provided in the KEGG database.

To rank the constructed regulatory loops in terms of their "significance," IntegraMiR applies a hypothesis testing procedure using Fisher's method [40]. The procedure employs Fisher's summary test statistic, given by Eq. (2.2) in Section 2.2.8, to combine the MHT-computed P values assigned to each node of the loop into one P value used as a ranking score for the entire loop. This does not apply to Type III loops, since these loops involve genes and not specific mRNA transcripts. Since the functional roles of regulatory loops are

different, IntegraMiR groups these loops into five distinct categories: Type I coherent FFL, Type I incoherent FFL, Type II coherent FFL, Type II incoherent FFL, and Type III loops – see Figs. 2.1 and 2.2. To provide additional flexibility in interpreting the results, IntegraMiR sorts Type II FFLs into two distinct subgroups, Type II-A and Type II-B, although this additional sorting may not be necessary. Within each group and subgroup, IntegraMiR ranks the deregulated loops by increasing scores, with lower scores corresponding to higher "significance," and highlights those loops discovered to be deregulated in a manner *consistent* with the underlying edge structure and the expression data, as determined by the rules depicted in Fig. 2.3 (see also Section 2.2.9). It moreover marks miRNA targets depending on whether these targets are predicted by the procedure or have been experimentally validated according to miRTarBase, or both. Note that "consistency" refers to the fact that the expression patterns of the nodes of a deregulated loop are in agreement with its regulatory edge structure. For example, a Type I coherent FFL is said to be consistently deregulated if it comprises an upregulated miRNA and downregulated TF and mRNA, or a downregulated miRNA and upregulated TF and mRNA; see Fig. 2.3.

To investigate the effectiveness of our proposed procedure, we apply IntegraMiR on mRNA/miRNA expression data from prostate tumor and normal samples. In the following, we discuss the details on the materials and techniques used in the development and evaluation of IntegraMiR.

2.2 Methods

2.2.1 Biological Samples

In this work, we use publicly available mRNA expression data obtained from a previously published study [13] involving normal and cancerous prostate tissue samples. The normal samples were acquired during radical prostatectomy from non-suspect (normal) peripheral areas of the prostate of 48 different individuals diagnosed with low-risk tumors. The cancerous samples were acquired from 47 patients diagnosed with high-risk tumors, before administering any medical treatment. Detailed discussion on the materials and methods used to obtain and prepare these samples can be found in [13]. We also use publicly available miRNA expression data from a previously published study [162] obtained from histologically confirmed *matched* malignant and peripheral nonmalignant prostate tissue samples extracted from 20 different patients with untreated prostate cancer (PCa). These samples were prepared from prostatectomy specimens using methods detailed in [162].

2.2.2 Expression Profiling

In [13], the total RNA extracted from each normal and cancerous prostate tissue sample was used to produce mRNA expression profiles for 17,324 human mRNAs. This was done by mRNA microarray hybridization using the Affymetrix (Santa Clara, CA) GeneChip Whole Transcript Sense Target Labeling Assay in conjunction with Affymetrix 1.0 Human Exon ST microarrays. The MIAME-compliant mRNA microarray data can be found in the

NCBI GEO database (www.ncbi.nlm.nih.gov/geo) with accession number GSE29079.

The tumor samples used to obtain the mRNA expression data were characterized by their disease subtype, based on their *TMPRSS2-ERG* gene fusion status, through a number of reliable assessments using *ERG* gene expression levels, nested RT-PCR, and quantitative PCR measurements [13, 18]. These data have also been validated with respect to an earlier study [150], which included matched miRNA expression data for a number of patients. Seventeen tumor samples were defined as *TMPRSS2-ERG* fusion-positive and twenty samples were defined as fusion-negative. The remaining ten tumor samples that could not be reliably categorized were labeled as unknown fusion status.

The miRNA profiling experiments performed in [162] used Affymetrix 1.0 GeneChip miRNA microarrays, whose content is derived from the miRBase miRNA database v11.0 (www.mirbase.org). These experiments produced expression data for 847 human miRNAs in *matched* normal and cancerous tissues. The data can be obtained from the NCBI GEO database using accession number GSE23022.

We should note here that several miRNA profiling studies have been published in the literature concerning PCa [4, 97, 111, 118, 153, 161]. However, results on deregulation of particular miRNA genes have been highly inconsistent [129]. Seeking support for the reliability of the miRNA expression data used in the present study, we should mention that a major factor that possibly contributes to these inconsistencies is known to be variations in the miRNA expression data due, for example, to a different proportion of stromal cells in tissue preparation. The previous miRNA microarray experiments are based on

micro-dissected tissue samples that avoid the previous issue. In addition, miRNA *in situ* hybridization experiments were run to evaluate the localization of miRNA-expressing cells and ensure that miRNA expression in tumor samples is indeed cancer cell-associated [162]. Moreover, the results were partially validated with RT-PCR and compared with a previous study on miRNA expression data from PCa tissue obtained by deep sequencing [147].

2.2.3 Data Preprocessing

IntegraMiR analyzes the raw mRNA and miRNA expression data using the statistical computing environment R (www.cran.r-project.org). Both types of data are background-corrected and normalized using quantile normalization [63]. In addition, the method employs the robust multi-array average (RMA) as a measure of mRNA and miRNA expression levels [63].

2.2.4 Multiple Hypothesis Testing/Surrogate Variable Analysis

Standard statistical tests used to identify differentially expressed genes between two conditions in a typical gene expression profiling study (as adopted by previous methods, e.g., see [17, 174]) become fundamentally flawed in the presence of unaccounted sources of variability (due to biological and experimental factors among others) [15, 86, 87]. As a consequence, many genes that are indeed differentially expressed in the data are not

detected, whereas many others are falsely declared as positives [16, 86].

To address this problem and effectively exploit the molecular subtyping information in the available mRNA expression data, IntegraMiR incorporates surrogate variable analysis (SVA) [87], together with multiple hypothesis testing (MHT), to identify differentially expressed genes between two conditions. The method uses the Bioconductor (www.bioconductor.org) package SVA (written in R) to perform SVA in order to take into account biological variabilities and batch effects due to molecular subtypes categorized by TMPRSS2-ERG gene fusion status in the tumor samples. This step has been shown to improve biological accuracy and reproducibility in genome-wide expression studies and enhances the quality of subsequent statistical analysis [86, 87].

IntegraMiR applies MHT to test for the null hypothesis $H_0^{(i)}$: $d_i = 0$ against the alternative hypothesis $H_A^{(i)}$: $d_i \neq 0$, where

$$d_i = \mu_i^{(t)} - \mu_i^{(n)}, \tag{2.1}$$

with $\mu_i^{(t)}$ and $\mu_i^{(n)}$ being the mean expression levels of the i-th transcript (mRNA or miRNA) in the tumorous and normal data, respectively. The Bioconductor package LIMMA (written in R), which implements a moderated t-statistic [137], is used on each data set to separately identify mRNAs and miRNAs that are differentially expressed between tumor and normal samples. Then, IntegraMiR applies the Benjamini-Hochberg method, described in [10], to control the false discovery rate (FDR) at 0.05. These steps produce two separate lists,

L_{mRNA} and L_{miRNA}, each containing 17,324 mRNAs and 847 miRNAs, with the corresponding FDR-adjusted P (or simply FDR) values and the direction of deregulation ($+1$ for upregulation and -1 for downregulation), as determined by the sign of the moderated t-statistic – see Table 2.1. The mRNAs and miRNAs with FDR ≤ 0.05 are considered as being differentially expressed between tumor and normal samples.

2.2.5 Gene Set Enrichment Analysis

To further evaluate the statistical significance of certain mRNA and miRNA transcripts deemed not to be differentially expressed by MHT, IntegraMiR uses LIMMA to perform gene set enrichment analysis (GSEA), taking into account known biological knowledge about these transcripts – see [142]. Specifically, by employing the molecular signatures database mSigDB v3.1 (www.broadinstitute.org/gsea/msigdb), the method uses GSEA to evaluate the significance of non-differentially expressed TFs in L_{mRNA} (MHT-based FDR > 0.05) for which the target gene sets can be obtained from mSigDB. IntegraMiR forms gene sets indexed by these TFs, with the elements of each gene set being those mRNAs in L_{mRNA} whose expressions are *directly* regulated by the indexing TF, as determined by mSigDB. It then uses GSEA to evaluate the statistical significance of each gene set to be enriched for deregulation, by using the default Wilcoxon rank-sum test. To adjust for multiple testing, IntegraMiR uses again the Benjamini-Hochberg method to control the FDR at 0.25 – see [142]. This step produces a list $L_{\text{TF}}^{\text{GS}}$ of TFs with the corresponding FDR values computed by MHT and GSEA – see Table 2.1. Only TFs with significantly

Table 2.1: Lists of mRNAs, TFs, miRNAs, and their targets used to construct deregulated loops and rank their statistical significance.

MHT	L_{mRNA}		L_{miRNA}
	◦ mRNAs ◦ p-values (MHT) ◦ direction of deregulation		◦ miRNAs ◦ p-values (MHT) ◦ direction of deregulation
GSEA	$L_{\text{TF}}^{\text{GS}}$	$L_{\text{miRNA}}^{\text{GS}}$	$L_{\text{KEGG}}^{\text{GS}}$
	◦ TFs from mSigDB in L_{mRNA} ◦ p-values (MHT) > 0.05 ◦ p-values (GSEA) ≤ 0.05 ◦ direction of deregulation	◦ miRNAs from miRTarBase in $L_{\text{miRNA}}^{\text{DeepSeq}}$ ◦ p-values (MHT) > 0.05 ◦ p-values (GSEA) ≤ 0.05 ◦ direction of deregulation	◦ KEGG pathways from mSigDB ◦ p-values (GSEA) ≤ 0.05
TARGET IDENTIFICATION	$L_{\text{mRNA}}^{\text{DiffExp}}$		$L_{\text{miRNA}}^{\text{DiffExp}}$
	◦ mRNAs ◦ p-values (MHT) ≤ 0.05 ◦ p-values (MHT) > 0.05 & p-values (GSEA) ≤ 0.05 ◦ direction of deregulation		◦ miRNAs ◦ p-values (MHT) ≤ 0.05 ◦ p-values (MHT) > 0.05 & p-values (GSEA) ≤ 0.05 ◦ direction of deregulation
	$L_{\text{TF}}^{\text{mRNA}}$	$L_{\text{TF}}^{\text{miRNA}}$	$L_{\text{miRNA}}^{\text{mRNA}}$
	◦ TFs in $L_{\text{mRNA}}^{\text{DiffExp}}$ ◦ mRNA targets from TRANSFAC in L_{mRNA} ◦ regulation type	◦ TFs in $L_{\text{mRNA}}^{\text{DiffExp}}$ ◦ transcriptional miRNA targets from TransmiR in $L_{\text{miRNA}}^{\text{DiffExp}}$ ◦ regulation type	◦ miRNAs in $L_{\text{miRNA}}^{\text{DiffExp}}$ ◦ predicted mRNA targets in L_{mRNA} using miRecords, amended with targets from miRTarBase

enriched gene sets (GSEA-based FDR ≤ 0.25) are included in this list. By combining lists L_{mRNA} and $L_{\text{TF}}^{\text{GS}}$, IntegraMiR finally forms a list $L_{\text{mRNA}}^{\text{DiffExp}}$ of mRNAs deemed to be differentially expressed by MHT or GSEA.

Likewise, IntegraMiR could use GSEA to further evaluate the statistical significance of non-differentially expressed miRNAs in L_{miRNA} for which it is able to obtain their targets from the experimentally verified database miRTarBase v3.5 (http://mirtarbase.mbc.nctu.edu.tw – see [58]). Unfortunately, the limited number of experimentally validated miRNA targets available in miRTarBase is a restricting factor in constructing appropriate and sufficiently large gene sets in order to reduce the resulting bias (e.g., due to small gene set size or experimental predilection – see [164] for a discussion on this issue). Due to bias and relatively small gene set sizes, GSEA produces an appreciable number of significantly enriched gene sets for miRNAs that are not detected to be differentially expressed by MHT (FDR > 0.05), a majority of which are expected to be false positives. A possible way to remedy this situation is to improve the statistical power of GSEA by constructing sufficiently large gene sets of miRNA targets that have been validated to be important in PCa by at least one reliable experimental procedure (see [164] for a discussion). For this reason, IntegraMiR limits this step to a list $L_{\text{miRNA}}^{\text{DeepSeq}}$ of 33 miRNAs that have been deemed to be significantly deregulated in PCa tissue using deep sequencing analysis [147]. Only gene sets having a minimum of *eight* elements are considered, as suggested in [134]. We should note here that it is not necessary to deal with this problem in the previous (and subsequent) application of GSEA, since all gene sets considered include a large and rather diverse number of elements in both

cases.

To proceed, IntegraMiR uses miRTarBase to form gene sets indexed by miRNAs in $L_{\text{miRNA}}^{\text{DeepSeq}}$, with MHT-based FDR values > 0.05 and with the elements of each gene set being the mRNA targets in L_{mRNA} of the indexing miRNA, as determined by miRTarBase. It then uses GSEA to evaluate the statistical significance of a particular gene set enriched for an inverse direction of deregulation with that of the miRNA. The reason IntegraMiR uses an inverse relation is because many experiments used in the past to identify miRNA targets, with their results recorded in miRTarbase, have traditionally focused on observing an inverse relation between the expression level of a miRNA and its experimentally validated target(s). This step produces a list $L_{\text{miRNA}}^{\text{GS}}$ of experimentally validated (by deep sequencing analysis) miRNAs with the corresponding FDR values computed by MHT and GSEA – see Table 2.1. Only miRNAs with significantly enriched gene sets (GSEA-based FDR ≤ 0.25) are included in this list. Finally, by combining lists L_{miRNA} and $L_{\text{miRNA}}^{\text{GS}}$, IntegraMiR forms a list $L_{\text{miRNA}}^{\text{DiffExp}}$ of miRNAs deemed to be differentially expressed by MHT or GSEA.

IntegraMiR also forms gene sets indexed by a specific KEGG signaling pathway included in mSigDB. The elements of each gene set are those mRNAs in L_{mRNA} determined by mSigDB to be in the indexing pathway. The method then uses GSEA to evaluate the statistical significance of a particular gene set to be enriched for deregulation in the available mRNA data. This step produces a list $L_{\text{KEGG}}^{\text{GS}}$ of gene sets, together with the associated KEGG signaling pathways and the corresponding GSEA-based FDR values – see Table 2.1. Only KEGG signaling pathways with significantly enriched gene sets (FDR ≤ 0.25) are included

in the list.

We should point out here that mSigDB provides miRNA target gene sets as well. However, using GSEA to evaluate the statistical significance of these gene sets to be enriched for deregulation produces poor results. We believe that this is due to the possibility that many miRNA target gene sets provided by mSigDB are false positives. Therefore, GSEA cannot produce meaningful statistical significance for these gene sets. As a consequence, IntegraMiR applies GSEA only on experimentally validated miRNA target gene sets in order to infer their statistical significance and complement the statistical analysis performed by simply using the available miRNA expression data.

2.2.6 Target Identification

Since the goal of IntegraMiR is to delineate deregulated miRNA/TF-mediated gene regulatory loops from evidence provided in available data, the method focuses on loops containing differentially expressed miRNAs and TFs (based on their individual expression levels – via MHT analysis – or through their target interactions – via GSEA analysis). Note however that the target mRNAs associated with the loops of interest may not necessarily be differentially expressed. This is due to the fact that differential expression of a TF may not imply differential expression of the targeting mRNA (a TF may produce insignificant regulation of transcription), whereas miRNA targeting may result in direct translational repression without affecting mRNA abundance. Moreover, simultaneous differential expression of the miRNA and TF nodes of an incoherent Type I or Type II FFL may result

in no deregulation of the associated mRNA since, in this case, downregulation (upregulation) of mRNA abundance by miRNA targeting may be counterbalanced by upregulation (downregulation) of transcription.

By following the previous rules, and for each differentially expressed TF in $L_{\text{mRNA}}^{\text{DiffExp}}$, IntegraMiR uses information available in TRANSFAC v7.0 (public) (www.gene-regulation.com/pub/databases.html – see [170]) to identify the *directly* regulated genes in L_{mRNA} as well as to determine the regulation type (activation or repression). To access this information and provide the input to IntegraMiR, we first obtained for each TF its TRANSFAC-compatible annotation using the automated sequence annotation pipeline (ASAP) system [76, 139]. This process yields a list $L_{\text{TF}}^{\text{mRNA}}$ containing differentially expressed TFs in $L_{\text{mRNA}}^{\text{DiffExp}}$, their gene targets in L_{mRNA}, and the regulation type (activation or repression) for each target gene – see Table 2.1. TFs not predicted to target any mRNAs in L_{mRNA} are not included in the list.

Likewise, IntegraMiR uses TransmiR v1.2 (http://202.38.126.151/hmdd/mirna/tf – see [163]) to identify differentially expressed TFs in $L_{\text{mRNA}}^{\text{DiffExp}}$ that *directly* regulate the transcription of miRNAs in $L_{\text{miRNA}}^{\text{DiffExp}}$. This produces a list $L_{\text{TF}}^{\text{miRNA}}$ containing TFs from $L_{\text{mRNA}}^{\text{DiffExp}}$, their corresponding transcriptional miRNA targets in $L_{\text{miRNA}}^{\text{DiffExp}}$, and the regulation type (activation or repression) for each targeted miRNA – see Table 2.1. TFs not predicted to target any miRNAs in $L_{\text{miRNA}}^{\text{DiffExp}}$ are not included in the list.

Finally, for each miRNA in $L_{\text{miRNA}}^{\text{DiffExp}}$, IntegraMiR performs sequence-based target prediction using miRecords (http://mirecords.umn.edu/miRecords – see [171]) with the filtering

parameter set equal to 2. As a consequence, targets for each miRNA are predicted by at least two (out of eleven) different sequence-based target prediction algorithms incorporated in miRecords. Moreover, for each differentially expressed miRNA with experimentally validated target information in miRTarBase, we identified those mRNA targets not predicted by miRecords. This produced a list $L_{\text{miRNA}}^{\text{mRNA}}$ of differentially expressed miRNAs in $L_{\text{miRNA}}^{\text{DiffExp}}$ with the corresponding sequence-based target predictions in L_{mRNA} amended with (experimentally validated) targets from miRTarBase – see Table 2.1. miRNAs not predicted to target any mRNAs in L_{mRNA} are not included in this list.

The reason we decided to use predictions by at least two different algorithms was to strike a balance between the number of false-positive and false-negative predictions. By setting the filtering parameter equal to 1, we obtain a large number of predictions (most of which are presumably false-positives) whereas by setting the filtering parameter equal to 3, we obtain a very small number of predictions (which presumably indicates a large number of false-negatives for the prediction). Note finally that miRecords provides a database for experimentally validated miRNA targets as well, but we decided to use miRTarBase instead, since the latter database is up-to-date, unlike the former which dates back to November 2010.

2.2.7 Construction of Regulatory Loops

IntegraMiR constructs Type I FFLs by first identifying (TF, mRNA) pairs using the list $L_{\text{TF}}^{\text{mRNA}}$. It then forms triplets (miRNA, TF, mRNA), such that a miRNA simultaneously

targets the TF and the mRNA, as determined by the list $L_{\text{miRNA}}^{\text{mRNA}}$ – see Fig. 2.1. Likewise, IntegraMiR constructs Type II FFLs by first identifying (TF, miRNA) pairs from the list $L_{\text{TF}}^{\text{miRNA}}$. It then forms triplets (miRNA, TF, mRNA), such that the mRNA is directly regulated by the TF and is simultaneously targeted by the miRNA, as determined by the lists $L_{\text{TF}}^{\text{mRNA}}$ and $L_{\text{miRNA}}^{\text{mRNA}}$, respectively. The method finally delineates all miRNA-target interactions in the four deregulated KEGG pathways under consideration (TGF-β Signaling, WNT Signaling, Prostate Cancer, and Adherens Junction) by first looking into the gene sets associated with each pathway (obtained from the KEGG database), by filtering out the genes that are not expressed in the data, and by identifying the targets of each miRNA as determined by the list $L_{\text{miRNA}}^{\text{mRNA}}$. In addition, IntegraMiR constructs Type III loops by taking gene pairs (G-1, G-2) such that their corresponding proteins could potentially interact with each other according to the pathway map provided by KEGG database. It then highlights triplets (miRNA, G-1, G-2) such that the miRNA is predicted to target at least one transcript from each gene, as determined by the list $L_{\text{miRNA}}^{\text{mRNA}}$. We carried out this step to identify, as an example, Type III loops in the KEGG Prostate Cancer Pathway for certain miRNAs.

Each edge depicted in Fig. 2.1 connecting a miRNA with its target is naturally repressing. The list $L_{\text{TF}}^{\text{mRNA}}$ provides the regulation type (activation or repression) for each edge connecting a TF with a mRNA whereas the list $L_{\text{TF}}^{\text{miRNA}}$ provides the regulation type (activation or repression) for each edge connecting a TF with a miRNA.

2.2.8 Significance Ranking of FFLs

For each constructed FFL, IntegraMiR calculates its statistical significance by employing the following procedure. First, by using the lists $L_{\text{mRNA}}^{\text{DiffExp}}$, $L_{\text{miRNA}}^{\text{DiffExp}}$, and L_{mRNA}, it associates with each node of the FFL a binary value (± 1), which indicates the direction of deregulation of the node. Moreover, it assigns the MHT-based FDR value corresponding to the particular transcript (TF, mRNA, or miRNA) represented by the node, which quantifies the significance of the transcript's deregulation. To evaluate the statistical significance of each FFL, IntegraMiR first assumes that the FFL is not deregulated if each one of its nodes $(1, 2, 3)$ is not deregulated. It then constructs a hypothesis testing procedure to test for the null hypothesis $H_0 : d_i = 0$, for every i, where $i = 1, 2, 3$, against the alternative hypothesis $H_A : d_i \neq 0$, for at least one i, where $i = 1, 2, 3$, with d_i given by Eq. (2.1), with $\mu_i^{(t)}$ and $\mu_i^{(n)}$ being the mean expression levels of the transcript (TF, mRNA, or miRNA) assigned at node i of the FFL in the tumorous and normal data, respectively. To do so, IntegraMiR uses Fisher's method [40, 169] based on the summary test statistic

$$T = -2 \ln(p_1 p_2 p_3) \,, \tag{2.2}$$

where p_1, p_2, and p_3 are the P values obtained by MHT for nodes 1, 2, and 3, respectively. Under the null hypothesis, each (non-adjusted) P value obtained by MHT will have a uniform distribution between 0 and 1. Assuming that these values are obtained from independent statistical tests, the statistic T follows a chi-square distribution with *six* degrees

of freedom, from which a combined value is obtained that is used to score each FFL.

We should note that these statistical tests depend on each other in general. It turns out that Fisher's method may result in a combined P value that will be smaller than the P value which could be obtained if dependencies among the statistical tests used could be appropriately taken into account. For this reason, we regard Fisher's method as producing a *score* for each FFL and not a formal P value [5]. As a consequence, we expect that IntegraMiR will produce a larger set of deregulated FFLs than a hypothesis testing method that properly considers the underlying dependence of the individual tests. In the absence of any prior information however, accounting for these dependencies is a difficult task [14, 77], which we cannot satisfactorily address in this dissertation.

2.2.9 Consistent Regulatory Loops

Since the functional roles of the FFLs considered in this dissertation are different, IntegraMiR groups them into five distinct categories: Type I coherent, Type I incoherent, Type II coherent, Type II incoherent, and Type III – see Fig. 2.1. In addition, the method sorts Type II FFLs into two distinct subgroups, Type II-A and Type II-B, and marks as "consistent" those loops discovered to be deregulated in a manner compatible with the underlying edge structure. To do so, note that molecular species joined by an activating edge are expected to exhibit *correlated* deregulation whereas species joined by a repressing edge are expected to exhibit *anti-correlated* deregulation. Taking this fact into account, IntegraMiR marks deregulated loops as being *consistent* by using the rules depicted in Fig. 2.3.

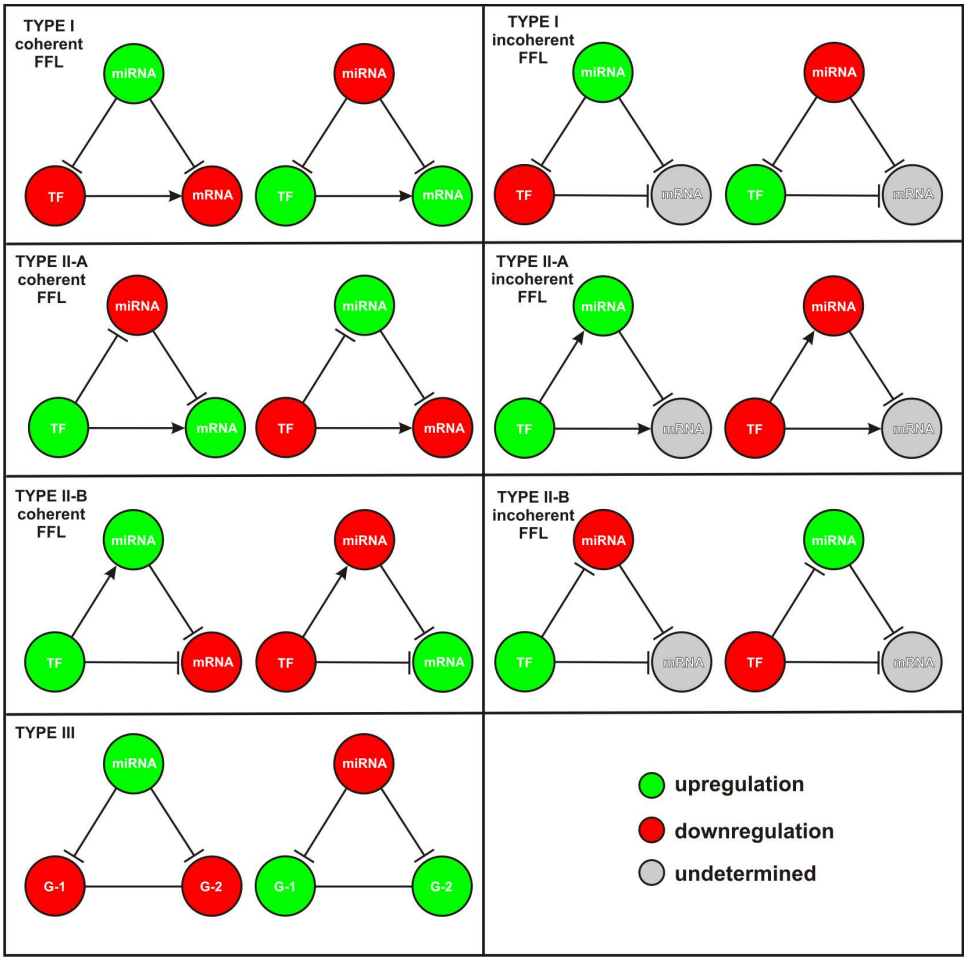

Figure 2.3: Consistency of deregulated loops. A deregulated loop is deemed to be *consistent* if the expression pattern of its nodes are in agreement with its regulatory edge structure. Any deregulated loop that does not satisfy this property is said to be *inconsistent*.

For example, a deregulated Type I coherent FFL is said to be consistent if it comprises an upregulated miRNA node and downregulated TF and mRNA nodes, or a downregulated miRNA node and upregulated TF and mRNA nodes. A deregulated FFL that does not follow these rules is said to be *inconsistent*.

2.2.10 Extracting Regulatory Loops

The results obtained by IntegraMiR are tabulated in the Supplementary Tables S5-S10 of [1] and contain a large number of deregulated Type I and Type II FFLs. To identify deregulated FFLs for specific miRNAs, TFs, or genes, we must search these results and extract those FFLs that contain the molecular species of interest. Moreover, identifying deregulated Type III loops for specific pairs of genes, requires construction of such loops from the results tabulated in Supplementary Table S11 of [1]. Extracting regulatory loops from the results can be done automatically.

2.3 Results

2.3.1 Identification of Extensive Transcriptional, Post-transcriptional and Signaling Deregulation in PCa

To investigate the effectiveness of IntegraMiR in delineating miRNA-mediated regulatory loops, we use mRNA microarray expression data, obtained from 48 normal and 47 prostate tumor tissue samples (NCBI GEO database, accession number GSE29079), as well as miRNA microarray expression data obtained from matched normal and cancerous tissue samples, extracted from 20 individuals (NCBI GEO database, accession number

GSE23022). For more information about this data, we refer the reader to Sections 2.2.1 and 2.2.2. After data preprocessing, IntegraMiR incorporates Surrogate Variable Analysis (SVA) [87], together with MHT, to identify differentially expressed genes between the two conditions. It has been shown that SVA increases the biological accuracy and reproducibility of analyses in genome-wide expression studies [86, 87]. IntegraMiR employs SVA to take into account biological variabilities due to molecular subtypes categorized by the status of *TMPRSS2-ERG* gene fusion, which has been identified in about half of all PCa cases and is a critical early event in the development and progression of this disease [29, 79, 152].

IntegraMiR first performs MHT, using a moderated t-statistic [137], to separately identify mRNAs and miRNAs that are differentially expressed between tumor and normal samples. This analysis identifies extensive transcriptional deregulation in the tumor tissue samples: 7,934 genes (out of 17,324) are found to be differentially expressed based on their statistical significance, with 164 of these genes being overexpressed by a fold change ≥ 2 or repressed by a fold change ≤ 0.5 – see Supplementary Tables S1 and S2 in [1]. The gene list we provide in Supplementary Table S2 contains important genes, such as *TARP*, *MYC*, *SNAI2 (SLUG)*, *WIF1* and *ERG* among others, which have been previously characterized in PCa.

Analysis of the corresponding miRNA expression data by MHT results in 18 (out of 847) differentially expressed human miRNAs, which we list in Table 2.2 (first 18 miRNAs) – see also the Supplementary Table S3 in [1]. Recently, deep sequencing analysis of miRNA expression profiles identified 33 miRNAs as being differentially expressed in PCa,

with miR-375, miR-200c, miR-143 and miR-145 exhibiting the most pronounced deregulation [147]. We compared the IntegraMiR results to the ones obtained by deep sequencing. Of the 18 miRNAs identified by IntegraMiR, 7 miRNAs (miR-200c, miR-20a, miR-375, miR-106a, let-7a, miR-21, and miR-106b) have been confirmed to be upregulated by deep sequencing analysis, whereas 2 miRNAs (miR-221 and miR-145) have been confirmed to be downregulated. The remaining 9 miRNAs identified by MHT were not detected by deep sequencing.

During the second step of IntegraMiR, application of GSEA on gene sets of TF targets obtained from mSigDB discovers 37 significantly deregulated TFs, which are not detected by the initial MHT step based on single gene analysis. A list these TFs can be found in Supplementary Table S4 of [1]. Interestingly, several of these TFs (e.g., NKX3-1, SMAD1, SMAD3, SRF, ETV4 and ELK1) are known to play important roles in PCa, as well as in other types of cancer.

Likewise, application of GSEA on gene sets of experimentally validated (by deep sequencing analysis) miRNA targets obtained from miRTarBase identifies 5 significantly downregulated miRNAs, which are not detected by MHT. We list these miRNAs in Table 2.2 (last 5 miRNAs). In both cases, and for each TF or miRNA, GSEA is performed based on the availability of gene sets in the data.

Finally, application of GSEA identifies 30 significantly deregulated signaling pathways, among the 186 KEGG signaling pathways available in mSigDB. We list the results in Table 2.3. Among other pathways, the list contains the TGF-β and WNT Signaling pathways,

Table 2.2: Differentially expressed miRNAs identified by IntegraMiR.

Rank	miRNA[1]	dir[2]	FDR (MHT)	FDR (GSEA)
1	miR-222	↓	6.58E-4	n/a
2	**miR-200c**	↑	1.32E-3	n/a
3	**miR-221**	↓	1.34E-3	n/a
4	**miR-20a**	↑	1.70E-3	n/a
5	miR-20b	↑	2.55E-3	n/a
6	miR-182	↑	3.52E-3	n/a
7	**miR-375**	↑	3.63E-3	n/a
8	miR-17	↑	4.12E-3	n/a
9	miR-93	↑	7.64E-3	n/a
10	**miR-145**	↓	9.58E-3	n/a
11	**miR-106a**	↑	1.04E-2	n/a
12	miR-141	↑	2.05E-2	n/a
13	mir-720	↑	2.27E-2	n/a
14	**let-7a**	↑	2.83E-2	n/a
15	miR-214	↓	2.85E-2	n/a
16	miR-200b	↑	2.95E-2	n/a
17	**miR-21**	↑	2.95E-2	n/a
18	**miR-106b**	↑	4.66E-2	n/a
19	**miR-125b**	↓	3.15E-1	9.02E-4
20	**miR-143**	↓	7.45E-1	1.06E-1
21	**miR-29a**	↓	8.62E-1	1.06E-1
22	**miR-24**	↓	8.79E-1	1.06E-1
23	**miR-199a**	↓	9.96E-1	1.06E-1

[1]Highlighted miRNAs have been confirmed by deep sequencing analysis.
[2]Direction of deregulation.

which have been implicated in PCa initiation and progression. Naturally, the results also include the Prostate Cancer and Adherens Junction pathways. The last pathway regulates intercellular adhesion that plays an important role in epithelial-to-mesenchymal transition (EMT), considered to be an important step in tumor progression [117, 172]. In the following, we limit our results and discussions to miRNA-target interactions associated with these four pathways.

Table 2.3: Significantly deregulated KEGG signaling pathways identified by IntegraMiR.

KEGG Signaling Pathway[1]	**FDR (GSEA)**
DILATED_CARDIOMYOPATHY	6.67E-4
ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC	6.67E-4
REGULATION_OF_ACTIN_CYTOSKELETON	8.34E-4
HYPERTROPHIC_CARDIOMYOPATHY_HCM	8.34E-4
TGF_BETA_SIGNALING_PATHWAY	3.68E-3
CALCIUM_SIGNALING_PATHWAY	4.09E-3
FOCAL_ADHESION	8.16E-3
ECM_RECEPTOR_INTERACTION	8.16E-3
WNT_SIGNALING_PATHWAY	8.77E-3
MAPK_SIGNALING_PATHWAY	1.40E-2
PROPANOATE_METABOLISM	1.55E-2
VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	1.76E-2
PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM	1.76E-2
FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	4.02E-2
PATHWAYS_IN_CANCER	4.36E-2
VASCULAR_SMOOTH_MUSCLE_CONTRACTION	4.36E-2
AXON_GUIDANCE	8.86E-2
UBIQUITIN_MEDIATED_PROTEOLYSIS	1.00E-1
MELANOGENESIS	1.00E-1
PROSTATE_CANCER	1.00E-1
ONE_CARBON_POOL_BY_FOLATE	1.20E-1
INOSITOL_PHOSPHATE_METABOLISM	1.49E-1
VASOPRESSIN_REGULATED_WATER_REABSORPTION	1.68E-1
ADHERENS_JUNCTION	1.71E-1
LONG_TERM_POTENTIATION	1.71E-1
PURINE_METABOLISM	1.71E-1
GLYCINE_SERINE_AND_THREONINE_METABOLISM	1.72E-1
GAP_JUNCTION	1.92E-1
ARGININE_AND_PROLINE_METABOLISM	2.32E-1
MELANOMA	2.50E-1

[1]Highlighted pathways used by IntegraMiR to construct Type III loops.

Lastly, and during the third and fourth steps, IntegraMiR constructs deregulated regulatory loops, sorts them into the seven groups depicted in Fig. 2.1 and ranks the Type I and Type II FFLs within each group using the scores computed by Fisher's summary test statistic. IntegraMiR predicts a large number of deregulated Type I and Type II FFLs, which can be found in Supplementary Tables S5-S10 of [1] (see also Fig. 2.4A): 2,104

Type I coherent, 649 Type I incoherent, 154 Type II-A coherent, 690 Type II-A incoherent, 486 Type II-B coherent, and 111 Type II-B incoherent. Moreover, the method predicts a large number of deregulated miRNA-target interactions that could potentially form Type III loops, which can be found in Supplementary Table S11 of [1]: 904 miRNA-mRNA pairs in the TGF-β Signaling Pathway, 1,611 miRNA-mRNA pairs in the WNT Signaling Pathway, 1,025 miRNA-mRNA pairs in the Prostate Cancer Pathway, and 896 miRNA-mRNA pairs in the Adherens Junction Pathway.

2.3.2 Discovery of Appreciable FFL-based Transcriptome Deregulation

To gain insight into the occurrence of deregulated Type I and Type II FFLs, we depict in Fig. 2.4A the fractions of deregulated FFL subtypes (among all deregulated FFLs predicted by IntegraMiR) grouped in terms of consistent and inconsistent deregulation (as defined in Section 2.2.9 and illustrated in Fig. 2.3) based on expression data. The results suggest that certain FFL subtypes contribute to a larger portion of the observed net FFL deregulation than other subtypes. Interestingly, consistent FFL deregulation accounts for about 35% of net FFL deregulation. This type of deregulation is important since its functional characteristics are corroborated by the available expression data, which provides a first level of evidence of their significance. For this reason, an experimentalist may want to first consider this type of FFL deregulation for validation. Among the consistently deregulated FFLs, the

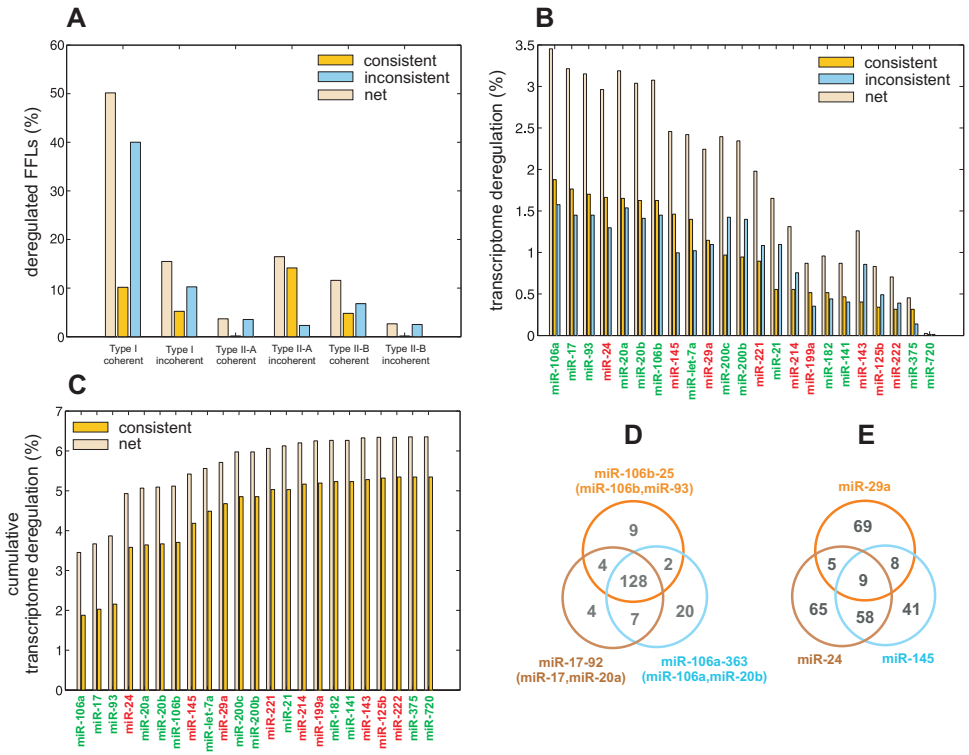

Figure 2.4: Predicted FFL-based transcriptome deregulation in PCa. (A) Distribution of the fraction of deregulated FFL subtypes grouped in terms of consistent and inconsistent deregulation based on expression data. (B) Percentages of transcriptome change due to significantly upregulated (in green) and downregulated (in red) miRNAs. (C) Cumulative percentages of transcriptome change due to significantly upregulated (in green) and downregulated (in red) miRNAs. (D) Venn diagram depicting the number of mRNA targets of six significantly upregulated miRNAs, miR-17 and miR-20a (from the miR-17/92 cluster), miR-106b and miR-93 (from the miR-106b/25 cluster), and miR-106a and miR-20b (from the miR-106a/363 cluster), which belong to the same family. (E) Venn diagram depicting the number of mRNA targets of three significantly downregulated tumor suppressor miRNAs, miR-24, miR-29a, and miR-145, which do not belong to one family.

Type II-A incoherent FFLs account for about 14% of net FFL deregulation, followed by Type I coherent FFLs, which account for 10%. On the other hand, Type I-A incoherent and Type II-B coherent FFLs each account for about 5% of net FFL deregulation, whereas, the two remaining subtypes, Type II-A coherent and Type II-B incoherent, account for less than 1%. It is striking however that 40% of FFL deregulation is attributed to inconsistent dereg-

ulation of Type I coherent FFLs. Inconsistent FFL deregulation suggests that the implied molecular interactions between the three nodes (miRNA, TF, mRNA) of a particular FFL may not be used to explain biological function on its own, based on the transcript levels of the nodes in the expression data. In this case, further investigation of underlying biological mechanisms that could affect the three FFL nodes is needed, including other FFLs sharing a node with the particular FFL under consideration.

To explain the previous result, note that we expect in the coherent case to observe a relatively smaller number of consistently than inconsistently deregulated FFLs since, for a coherent FFL to be consistently deregulated, the abundance of the three associated molecular species (miRNA, TF, and mRNA) must satisfy the rules depicted in Fig. 2.3 (see also Section 2.2.9). The required conditions however may not be observed in the data, since the abundance of a molecular species may be influenced by several FFLs or by factors other than FFL regulation. Clearly, the results depicted in Fig. 2.4A corroborate this remark. On the other hand, IntegraMiR predicts that Type I coherent FFL deregulation accounts for an appreciable portion (50%) of net FFL deregulation which, together with the previous remark, explains the high percentage (40%) of net FFL deregulation due to inconsistently deregulated Type I coherent FFLs.

By examining the constituent interactions that form deregulated FFLs, we determined, for each significantly deregulated miRNA, the percentage of transcriptome deregulation attributed to that miRNA. The results are depicted in Fig. 2.4B, ranked in terms of decreasing percentages of consistent deregulation. We call a miRNA-target interaction to be

consistent, if the miRNA and the associated mRNA target exhibit anti-correlated deregulation in the data. Note that miR-106a is responsible for the most consistent (1.88%) and the most inconsistent (3.45%) transcriptome deregulation, whereas miR-720 has negligible transcriptome changes associated with it. Finally, the cumulative distributions depicted in Fig. 2.4C reveal that 6.35% of transcriptome changes between normal and cancer samples are due to FFLs with significantly deregulated miRNA nodes, with 5.34% of the changes being accounted for by consistently deregulated miRNA-target interactions.

2.3.3 Consonancy with MiRNA Family Co-targeting Hypothesis

Among the top miRNAs depicted in Fig. 2.4B are members of three miRNA clusters that have been investigated in other types of cancers as well [104]: miR-17/92 on human chromosome 13 (with genomic locus 13q31.3) and its two cluster paralogs, miR-106a/363 on chromosome X (Xq26.2) and miR-106b/25 on chromosome 7 (7q22.1). Members of these clusters have been established to play essential roles in the normal development of heart, lungs, and the immune system and are involved in tumor formation with oncogenic roles [102, 115, 159]. More importantly, miR-17 and miR-20a (from the miR-17/92 cluster), miR-106a and miR-20b (from the miR-106a/363 cluster), as well as miR-106b and miR-93 (from the miR-106b/25 cluster) belong to the same family of miRNAs (i.e., miRNAs with identical seed regions) and are deemed to be significantly upregulated

by IntegraMiR. Note however that individual miRNAs on the same cluster could exhibit varied levels of expression and, for some miRNAs, no expression at all in certain cell lines [52, 136]. Along these lines, several miRNAs in the miR-17/92 cluster and its two paralogs (in particular, miR-18a, miR-19a, miR-19b-1 and miR-92a-1 from the miR-17/92 cluster, miR-18b, miR-19b-2, miR-92a-2 and miR-363 from the miR-106b/25 cluster, as well as miR-25 from the miR-106a/363 cluster) are not identified as being differentially expressed based on the expression data we used in this study.

Recent work suggests that members of the same family of miRNAs tend to target common transcripts due to their shared seed sequences [156]. The results obtained by IntegraMiR corroborate this hypothesis. In Fig. 2.4D, we use a Venn diagram to depict the numbers of mRNA targets predicted by IntegraMiR for the previous six miRNAs (obtained from miRNA-target interactions among all FFLs in our results – see Supplementary Tables S5-S10 in [1]). Clearly, a high level of overlap exists among the three target sets. In particular, our results predict that all six miRNAs target a set of 128 different mRNAs. This finding has also been observed by using an alternative method and different data sets [43], suggesting that cooperation among the six deregulated miRNAs may be present in other cancer types as well.

On the other hand, the top three miRNAs miR-24, miR-29a, and miR-145 in Fig. 2.4B which were found by IntegraMiR to be significantly downregulated, do not belong to one family and are not known to reside on a common cluster according to the miRBase (www.mirbase.org) database. The results depicted in Fig. 2.4E show that, in this case, the

amount of overlap is less pronounced than the one depicted in Fig. 2.4D. It is important to note that these three miRNAs have been hypothesized to possess tumor suppressor roles: miR-24 has recently been shown to suppress expression of two crucial cell cycle control genes, *E2F2* and *MYC* [81], low levels of miR-29a have been attributed to the methylation of its promoter region in PCa [90], and miR-145 is hypothesized to play roles in several types of cancer [125].

2.3.4 Discovery of Appreciable FFL-based MiRNA-TF Co-regulation

We now focus our attention on FFL-based miRNA-TF co-regulation. In Fig. 2.5A, we depict the numbers of coherent and incoherent deregulated FFLs predicted by IntegraMiR for each type of miRNA-TF interaction whereas, in Fig. 2.5B, we depict the percentages of consistently and inconsistently deregulated miRNA-TF interactions under each category. The results suggest that, in PCa, both coherent and incoherent FFLs are deregulated, although the total coherent FFLs outnumber the incoherent ones, an observation that is especially true when the miRNA represses the TF (Type I). Moreover, the most prevalent FFL deregulation involves repression of the TF by the miRNA (Type I coherent and incoherent), followed by FFL deregulation that involves activation of the miRNA by the TF (Type II-A incoherent and Type II-B coherent). On the other hand, deregulation of FFLs that involve repression of the miRNA by the TF (Type II-A coherent and Type II-B incoherent)

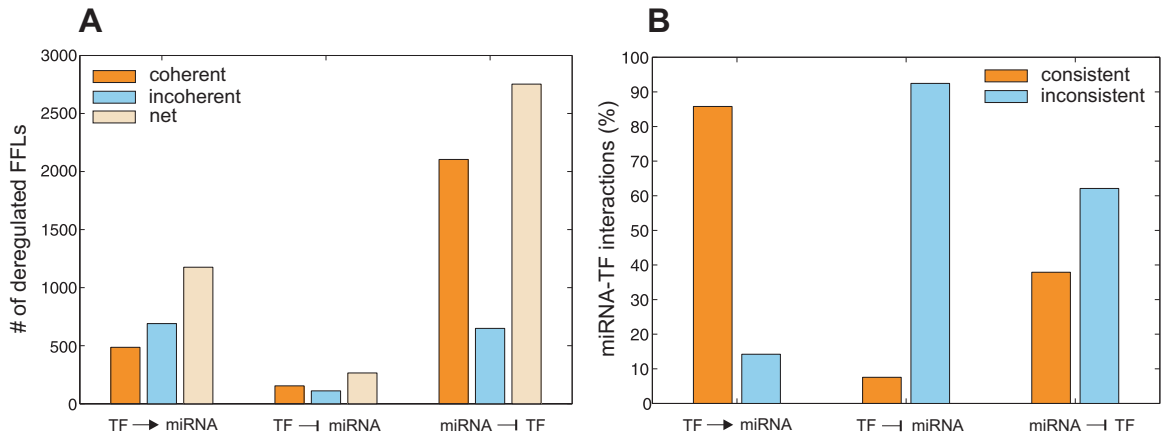

Figure 2.5: Predicted FFL-based miRNA-TF co-regulation. (A) Numbers of coherent and incoherent deregulated FFLs for each type of miRNA-TF interaction. (B) Percentages of consistently and inconsistently deregulated FFLs under each miRNA-TF interaction type depicted in (A).

is not substantial. Note also that consistent deregulation of FFLs that involve activation of the miRNA by the TF (Type II-A incoherent and Type II-B coherent) is appreciably more prevalent than inconsistent deregulation whereas the opposite is true for the case of FFLs in which the TF represses the miRNA.

All miRNA-TF pairs associated with the deregulated FFLs predicted by IntegraMiR (obtained from miRNA-TF interactions among all the FFLs in our results) are listed in Supplementary Table S12 of [1], categorized by their interaction type. As a notable example, the six miRNAs considered in Fig. 2.4D appear in the list as being consistently deregulated together with the *MYC* oncogene, which acts as their transcriptional activator. We investigated how many of the 128 common mRNAs targeted by these six miRNAs were predicted to form FFLs with *MYC*. IntegraMiR predicts 79 of the 128 mRNAs to be under the regulatory control of *MYC*, divided into two sets, with 33 mRNAs being in the first set and 46 mRNAs in the second. All six miRNAs interact with the first set of mRNAs in

Type II-B coherent FFL configuration and with the second set in Type II-B incoherent FFL configuration. Among these mRNAs, *APP* from the first set and *E2F1* from the second set have experimentally validated interactions with these miRNAs according to miRTarBase.

2.3.5 Discovery of Bona Fide MiRNA-mediated Regulatory Networks

To demonstrate the significance of the results obtained by IntegraMiR from a mechanistic point of view, we focus on two biological settings known to play crucial roles in PCa and other types of cancer. This will help us explain the functional roles of regulatory modules and illustrate how one can use these modules to build an integrated network model for a specific biological setting or molecular species of interest.

2.3.5.1 TP53 miRNA-mediated apoptotic network

We first consider the miR-125 family of miRNAs, which is highly conserved throughout diverse species from nematodes to humans. Members of this family, such as miR-125a, miR-125b, and miR-125b-2, have been validated to be downregulated, exhibiting disease-suppressing properties in many conditions as well as disease-promoting functions [146]. It turns out that miR-125b is identified by IntegraMiR to be significantly downregulated – see Table 2.2. It has been recently suggested that miR-125b is an important component of a TP53 (p53) tumor-suppressor network whereas significant negative correlation has

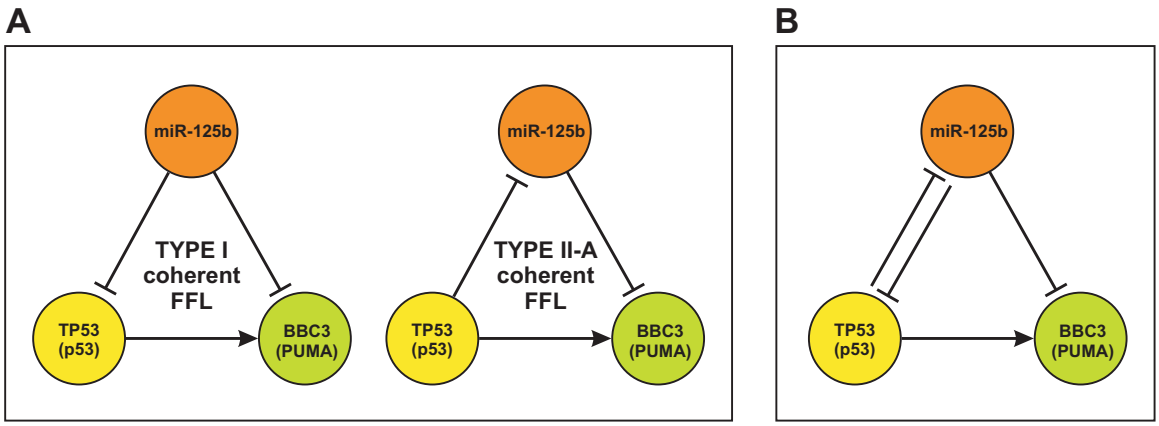

Figure 2.6: TP53 miRNA-mediated network model for apoptosis. IntegraMiR identifies two deregulated FFLs in PCa that model regulatory interactions among miR-125b, TP53 (p53), and BBC3 (PUMA). (A) Type I coherent and Type II-A coherent FFLs. (B) TP53 miRNA-mediated network model for apoptosis obtained by combining the two FFLs in (A).

been reported between miR-125b and TP53 [11, 78]. Moreover, it has been shown that the p53-upregulated modulator of apoptosis *BBC3 (PUMA)* and *NOXA* are direct targets in p53-mediated apoptosis localized to mitochondria [112].

To investigate systemic relations among these molecules of interest, we identified all deregulated FFLs predicted by IntegraMiR that contain miR-125b, TP53 (p53), BBC3 (PUMA) and NOXA. To focus our discussion on highly relevant FFLs, we consider only FFLs with nodes comprised of one of the four species of interest. We could not find FFLs that contain NOXA. However, we found one Type I coherent FFL and one Type II-A coherent FFL comprised of the other three species – see Fig. 2.6A. Both FFLs are deemed by IntegraMiR to be deregulated in the prostate expression data.

The Type I coherent FFL suggests that miR-125b represses *BBC3* while it reinforces this repression by targeting its transcriptional activator TP53. The Type II-A coherent FFL suggests that TP53 induces the transcription of *BBC3* while it reinforces this induction by

repressing miR-125b, an inhibitor of *BBC3*.

From a systemic point of view, if the Type I coherent FFL is functional in a specific condition in which miR-125b is significantly upregulated, we would expect the expressions of both *TP53* and *BBC3* to be repressed. As a consequence, miR-125b would assume an anti-apoptotic role in this setting. A similar argument can be made when miR-125b is significantly downregulated. As for the Type II-A coherent FFL, if TP53 is upregulated and active as a TF, we would expect miR-125b to be downregulated. As a consequence, BBC3 is expected to be significantly upregulated due to the concurrent upregulation of its transcriptional inducer, TP53, and the repression of its inhibitor, miR-125b. It is noteworthy that one cannot always expect to observe these exact relations in mRNA/miRNA expression data. It turns out that both FFLs depicted in Fig. 2.6A are deregulated inconsistently based on the expression data.

The previous steps provide a fundamental understanding of the underlying structure of TP53 miRNA-mediated apoptotic network, which may not be directly attainable by looking at individual molecular interactions. In particular, by combining the two FFLs depicted in Fig. 2.6A, we obtain the simple network depicted in Fig. 2.6B. This network accentuates the mutual inhibition between miR-125b and the pro-apoptotic interaction between TP53 and BBC3, which is in line with the earlier reported observation of significant negative correlation between miR-125b and TP53 [11, 78]. The underlying double negative feedback means that upregulation of miR-125b will inhibit TP53 which will derepress miR-125b, a situation that can lead to the repression of *BBC3*. On the other hand, downregulation

of miR-125b will derepress TP53 which will further repress miR-125b, a situation that may lead to significant activation of *BBC3* and thus apoptosis. Double negative feedback loops are known to act as toggle switches that lead to different cell fates [2]. Interestingly, both *TP53* and *BBC3* have been validated to be targets of miR-125b according to miRTarBase. Moreover, the Type I FFL discussed above has been recently reported in [135], thus demonstrating the validity of the previous IntegraMiR predictions.

2.3.5.2 MYC-E2F1 miRNA-mediated cell proliferation network

It is well known that deregulated expression or malfunction of the transcription factor MYC is one of the most common abnormalities in human cancers. Moreover, E2F1 is a member of the E2F family of TFs which are critical regulators of cell cycle and apoptosis. This TF regulates *MYC* and is transcriptionally targeted by MYC. Considering the fact that the miR-17/92 cluster and its paralogs have recently been shown to be tightly linked to the functions of MYC and E2F1 in the regulatory circuitry that controls cell proliferation [104, 108, 109, 159], we decided to identify all miRNA regulators predicted by IntegraMiR to interact with these critical TFs. This allowed us to delineate the regulatory network depicted in Fig. 2.7, which we constructed from 18 distinct FFLs: 8 Type I coherent, 2 Type II-A coherent, and 8 Type II-A incoherent. A total of 9 miRNAs were predicted to interact both with *MYC* and *E2F1*, with 8 of the miRNA-target interactions being identified by the predictive module of IntegraMiR as being true-positives, 2 being identified as false-negatives, and 3 being novel predictions that need to be experimentally validated.

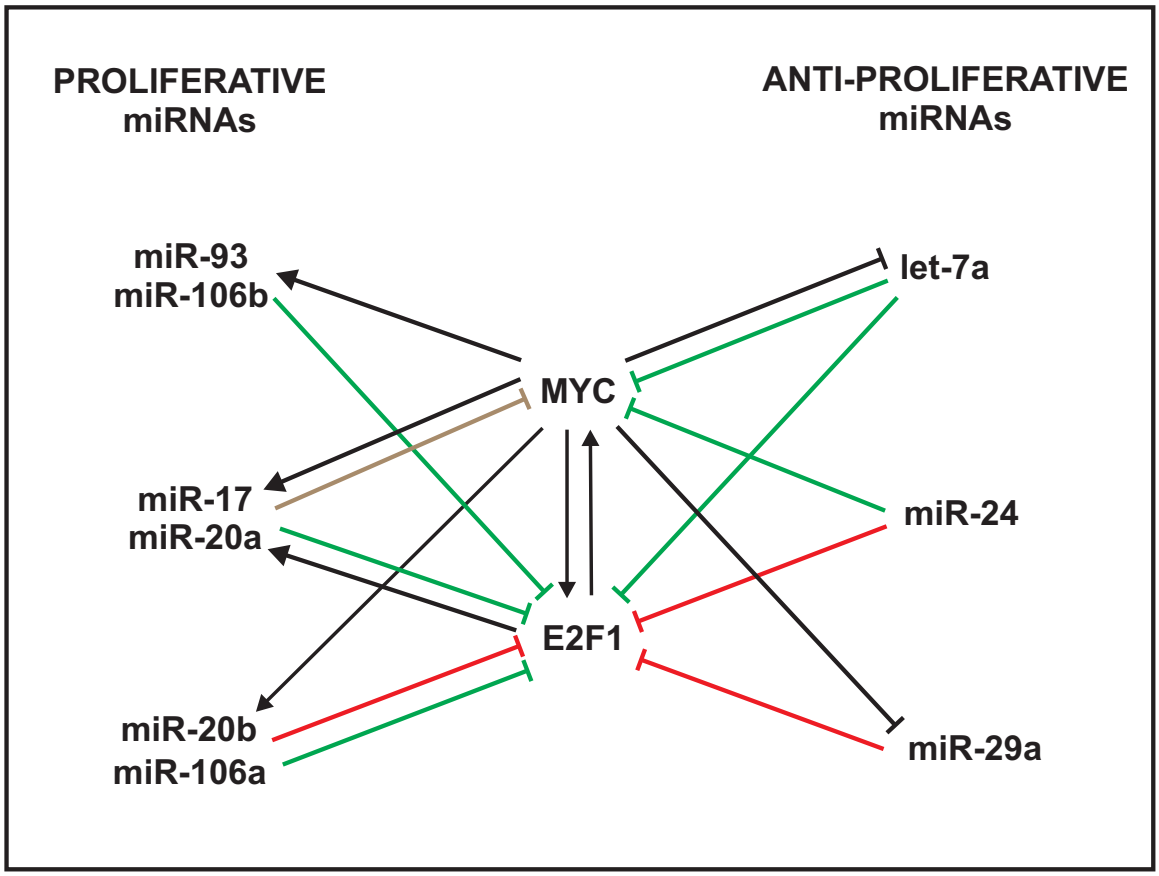

Figure 2.7: MYC-E2F1 miRNA-mediated network model for cell proliferation. A network of proliferative and anti-proliferative miRNAs interacting with MYC and E2F1 predicted by IntegraMiR. This network consists of 18 distinct FFLs: 8 Type I coherent, 2 Type II-A coherent, and 8 Type II-A incoherent. Green edges depict true-positive miRNA-target interactions identified by the predictive module of IntegraMiR, the brown edge predicts a false-negative miRNA-target interaction, and the red edges depict novel miRNA-target interactions.

From a mechanistic point of view, the negative feedback loops and incoherent FFLs on the left-hand-side of Fig. 2.7 ensure a tightly controlled regulation of cell proliferation. It has been argued in [102] that high levels of E2F proteins, especially E2F1, can induce apoptosis, and the negative feedback with miR-17 and miR-20a may dampen E2F activity following a physiologic proliferative signal, thereby promoting cell division rather than cellular death. On the other hand, the double-negative feedback loops and coherent FFLs on

the right-hand-side of Fig. 2.7 suggest anti-proliferative roles for the corresponding miRNAs, since these interactions repress MYC/E2F1 induced proliferation. As we mentioned before in our discussion related to Fig. 2.4E, miR-24 and miR-29a exhibit tumor-suppressor roles, which is compatible with the network depicted in Fig. 2.7. The miRNA let-7a has also been given a tumor-suppressor role in PCa [35], as well as in lung and renal cancers [95, 110].

2.3.6 Tumor-suppressor Roles for MiR-24, MiR-29a and MiR-145 in PCa

IntegraMiR identifies a large number of deregulated miRNA-target interactions in the four pathways we consider in this dissertation: 906 interactions in the TFG-β Signaling Pathway, 1,610 interactions in the WNT Signaling Pathway, 1,017 interactions in the Prostate Cancer Pathway, and 895 interactions in the Adherens Junction Pathway – see Supplementary Table S11 in [1]. These pairs can potentially be used to form Type III regulatory loops.

To illustrate the functional scope and relevance of these interactions, we focus on the top three miRNAs depicted in Fig. 2.4B found by IntegraMiR to be significantly downregulated. These are the tumor suppressor miRNAs miR-24, miR-29a, and miR-145 studied in Fig. 2.4E. Using these miRNAs, we considered the deregulated miRNA-target interactions predicted by IntegraMiR and identified, as an example, those interactions relevant to the

KEGG Prostate Cancer Pathway. IntegraMiR predicts a considerable number of deregulated interactions (45 for miR-24, 41 for miR-29a, and 40 for miR-145) with many common targets in this pathway. This may further be used to support the collaborative, tumor-suppressor role of these miRNAs in PCa, despite the fact that their predicted, genome-wide co-targeting features are relatively not much pronounced – see Fig. 2.4E.

By using Supplementary Table S11 in [1], we also identified the consistently deregulated Type III regulatory loops associated with the three miRNAs, miR-24, miR-29a, and miR-145, in the KEGG Prostate Cancer Pathway by excluding the missing pathway interactions as well as interactions with indirect effects, as defined by the KEGG database [69, 70]. We depict the results in Fig. 2.8. From all predicted interactions, only the interaction between miR-145 and IGF1R, a product of the *GFR* gene, as well as the interaction between miR-29a and PIK3R1, a product of the *PI3K*, are known (i.e., are true-positives). It turns out that several Type III loops predicted by IntegraMiR encompass genes that have established oncogenic roles, such as the genes in the PI3K-Akt backbone and the *RAS* and *RAF* genes in the MAPK signaling section of the pathway. This observation thus provides further support for the tumor-suppressor roles of these miRNAs in PCa.

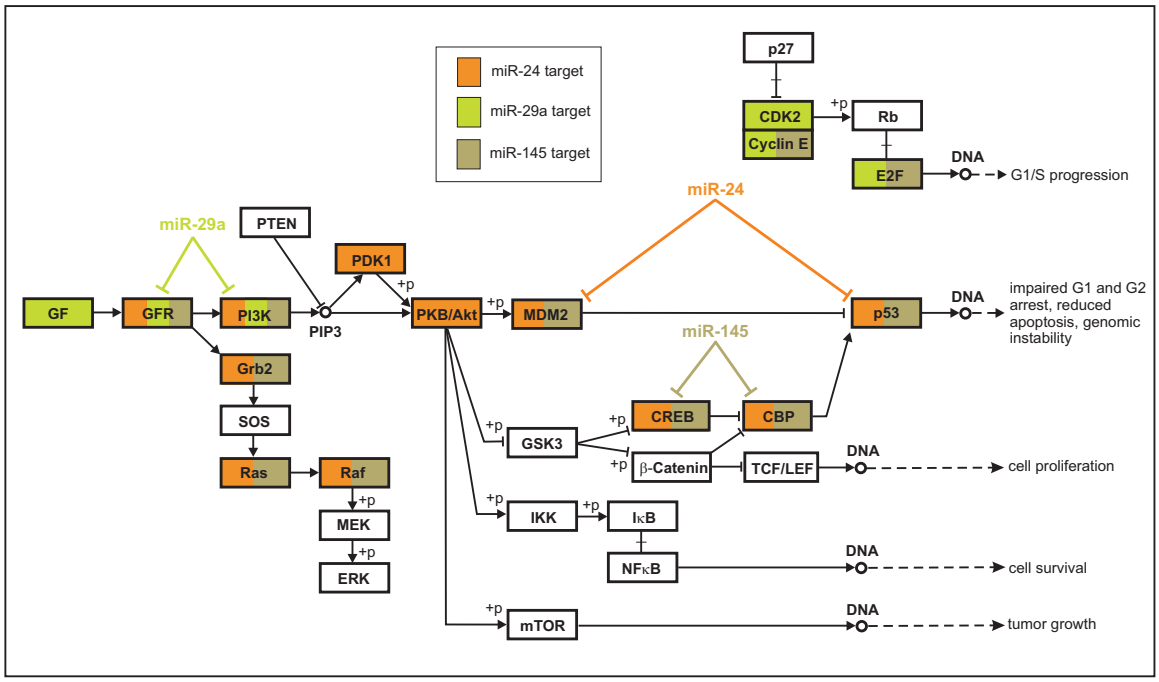

Figure 2.8: Predicted deregulated Type III regulatory loops in the Prostate Cancer Pathway. Portion of the Prostate Cancer Pathway, adopted from the KEGG database, with the targets of miR-24, miR-29a and miR-145 that participate in deregulated Type III loops being color-coded. One example of a deregulated Type III loop is shown for each miRNA. All depicted Type III loops are consistent, in the sense that the corresponding miRNA-target interactions are anti-correlated according to the data.

2.3.7 A Novel Regulatory Circuit for Epithelial-to-Mesenchymal Transition (EMT)

EMT is a complex gene expression program characterized by loss of cell adhesion through repression of *CDH1* (E-cadherin) and activation of genes associated with motility, invasion and stemness [28]. EMT is activated during embryonic development and adult tissue remodeling. In epithelium-derived tumors however, EMT seizes to promote metastasis and gain of stem cell phenotypes [117]. Since modulation of *CDH1* expression levels is considered to be a major theme of epithelial plasticity, both in non-oncogenic and onco-

genic EMT, we sought to construct and investigate an integrated circuit that controls EMT in PCa based on IntegraMiR predictions.

A natural approach towards this goal is to first identify the most relevant molecular species to build an initial network and subsequently expand this network with additional species. Since our main interest here is to determine FFLs mainly involved in pathological conditions related to EMT and since the most common biochemical change associated with EMT is loss of *CDH1* expression, we decided to focus on *CDH1* repressors and their corresponding regulatory network. *CDH1* transcriptional repressors, such as SNAI1 (SNAIL), SNAI2 (SLUG), ZEB1, ZEB2 (SIP1), E12/E47, and TWIST have traditionally been implicated in promoting EMT in various systems of embryonic development and tumor progression [28, 100]. Among these repressors, we found that SNAI2 and ZEB1 are associated with FFLs predicted by IntegraMiR – see Supplementary Tables S5-S10 in [1]. It is important to note that the TGF-β Signaling Pathway induces the transcription of *SNAI2 (SLUG)*, which in turn activates *ZEB1* [103, 167]. Furthermore, the miR-200 family of miRNAs (miR-200a, miR-200b, miR-200c, miR-141 and miR-429) has been shown to play a major role in EMT [28, 48]. Among the family members, miR-200b, miR-200c and miR-141 have been identified by IntegraMiR to be significantly deregulated in PCa – see Table 2.2.

To delineate a basic network for EMT regulation, we first single out all deregulated FFLs whose nodes comprise only entries among the molecular species we have identified: miR-200b, miR-200c, miR-141, CDH1, SNAI2, and ZEB1. These FFLs are deemed to be consistently deregulated by IntegraMiR. For miR-141, we discovered two loops whereas

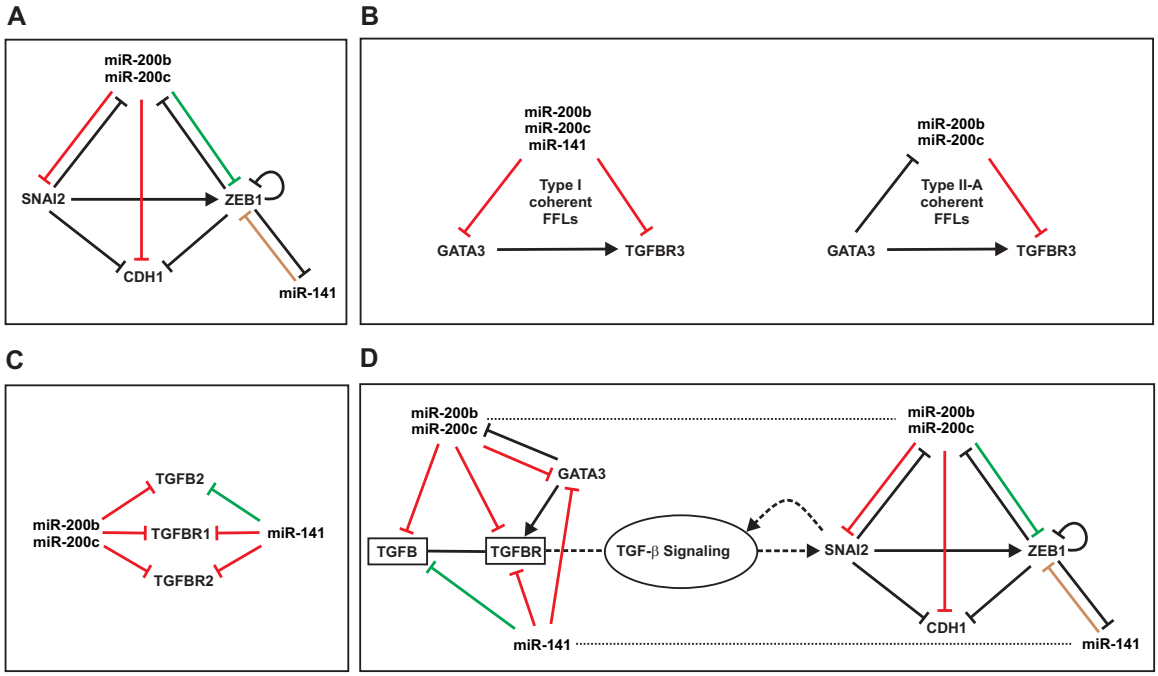

Figure 2.9: Predicted regulatory circuits controlling EMT. (A) An initial regulatory circuit, predicted by IntegraMiR, controlling EMT in PCa through regulation of CDH1 (E-cadherin) transcriptional repressors. This network consists of 14 distinct FFLs: 2 Type I coherent, 5 Type I incoherent, 2 Type II-A coherent, and 5 Type II-B incoherent. (B) The five FFLs predicted to be (consistently) deregulated in PCa by IntegraMiR comprising miR-200b, miR-200c, or miR-141, and GATA3 and TGFBR3. (C) The nine deregulated miRNA-target interactions involving miR-200b, miR-200c, and miR-141 as well as the TGFB ligands and receptors. (D) An extended integrated regulatory circuit, predicted by IntegraMiR, controlling EMT through TGF-β signaling and regulation of CDH1 transcriptional repressors. In these figures, green edges depict true-positive miRNA-target interactions identified by the predictive module of IntegraMiR, brown edges represent false-negative miRNA-target interactions, whereas red edges depict novel miRNA-target interactions.

for miR-200b and miR-200c, we discovered six loops for each miRNA with identical types.

We then constructed the network depicted in Fig. 2.9A by combining these FFLs.

To extend this basic network, we regard the fact that TGF-β signaling induces the transcription of *SNAI2* and consider the recent discovery that SNAI2 and TGF-β signaling interact in a positive feedback loop [30, 100]. We then hypothesized that we may observe

a (mutually) inhibitory relation between members of the miR-200 family and upstream factors in TGF-β signaling due to the fact that these miRNAs interact with SNAI2 in a mutually inhibitory fashion, as predicted by the network depicted in Fig. 2.9A. To constrain this investigation to a tractable number of transcripts, among the numerous transcripts associated with TGF-β signaling, we focus on the very first elements of this pathway: three TGFB isoforms (TGFB1, TGFB2, TGFB3) and three TGFB receptors (TGFBR1, TGFBR2, TGFBR3).

We should note here that, among TGFB cell surface receptors, *TGFBR3* has the most abundant expression and it shows the highest affinity for binding TGFB2 ligand among all three TGFB ligand isoforms. While TGFBR3 does not have a functional kinase domain to activate TGF-β signaling, it helps TGFB ligands be presented to TGFBR2, which leads to the association and phosphorylation of TGFBR1 and subsequent activation of TGF-β signaling by phosphorylation of SMAD2 or SMAD3 proteins [36]. Reduced or loss of *TGFBR3* expression has been observed in many types of cancer, such as prostate, pancreatic, breast, renal, and lung cancer [25, 34, 39, 47, 157].

We identified all FFLs predicted to be deregulated by IntegraMiR (see Supplementary Tables S5-S10 in [1]) comprising miR-200b, miR-200c, or miR-141, and one of the TGFB ligand isoforms or one of the TGFB receptors. This produced the three Type I coherent and the two Type II-A incoherent FFLs depicted in Fig. 2.9B all of which are deemed to be consistently deregulated in the data. We also identified all deregulated miRNA-target interactions for miR-200b, miR-200c, and miR-141 associated with the KEGG TGF-β

Signalling Pathway (see Supplementary Table S11 in [1]), with the target being one of the TGFB ligand isoforms or one of the TGFB receptors. We depict the results in Fig. 2.9C, which shows that each of these miRNAs targets *TGFB2*, *TGFBR1*, and *TGFBR2*. Among these interactions, only *TGFB2* has been experimentally verified to be a target of miR-141 according to miRTarBase.

By incorporating the results depicted in Figs. 2.9A-C, we obtain the extended circuit for EMT regulation depicted in Fig. 2.9D. To simplify presentation, we lump specific interactions of the miRNAs with individual TGFB receptors in a single block. Interestingly, this circuit predicts a mutually inhibitory relation between miR-200b, miR-200c and GATA3, a recently discovered transcriptional activator for *TGFBR3* [24]. Moreover, miR-200b, miR-200c, and miR-141 are predicted to repress the upstream TGFB2 ligand and receptors in a Type III regulatory loop. The resulting integrated regulatory circuit provides a hypothesis for a novel and more comprehensive model for regulation of EMT at the transcriptional, post-transcriptional and signaling levels, by means of miR-200 family members, TGF-β signaling and the corresponding transcriptional program.

2.3.8 A Relatively Comprehensive Model for PCa Development

To discern the effectiveness of the integrative analyses we carry out in this study, we combined information from the results depicted in Fig. 2.7 and Fig. 2.8, as well as cur-

rent knowledge of certain crucial genetic and epigenetic alterations in PCa (which we will be discussing shortly), to delineate the model depicted in Fig. 2.10. This model encapsulates some major sources of deregulation in PCa at the transcriptional, post-transcriptional, signaling, and genetic/epigenetic levels, as opposed to models that only consider deregulation at just one level, which may not be capable of capturing the overall behavior of the underlying network. We use this model to discuss how genetic and epigenetic alterations could propagate in cellular regulatory networks through circuits identified in this study and, therefore, adversely affect gene regulation. These pieces of crucial information represent a relatively comprehensive model for PCa development.

It has been demonstrated that chromosomal translocation involving *TMPRSS2* (PSA-regulated gene transmembrane protease, serine 2), an androgen receptor (AR)-regulated gene, and a member of the ETS family of TFs (predominantly ERG) is present in about half of all PCa cases [152]. This rearrangement in prostate cancer leads to androgenic induction of *ERG* expression (see Fig. 2.10) and the critical outcomes associated with its overexpression in PCa [151]. In particular, it has been suggested that *ERG* overexpression in PCa may contribute to the neoplastic process by activating *MYC* and by abrogating prostate epithelial differentiation [144]. Moreover, global analysis of copy-number alterations (CNAs) in PCa has reported dramatic amplifications of oncogenes, such as *MYC* (on 8q24.21) and *AR* (Xq12), deletions of tumor suppressor genes, such as *PTEN* (10q23.31), *RB1* (13q14.2), *TP53* (17p31.1) and *CDKN1B* (due to the broader deletion of the 12p13.31-p12.3 genomic region), and interstitial 21q22.2-3 deletion spanning *ERG* and

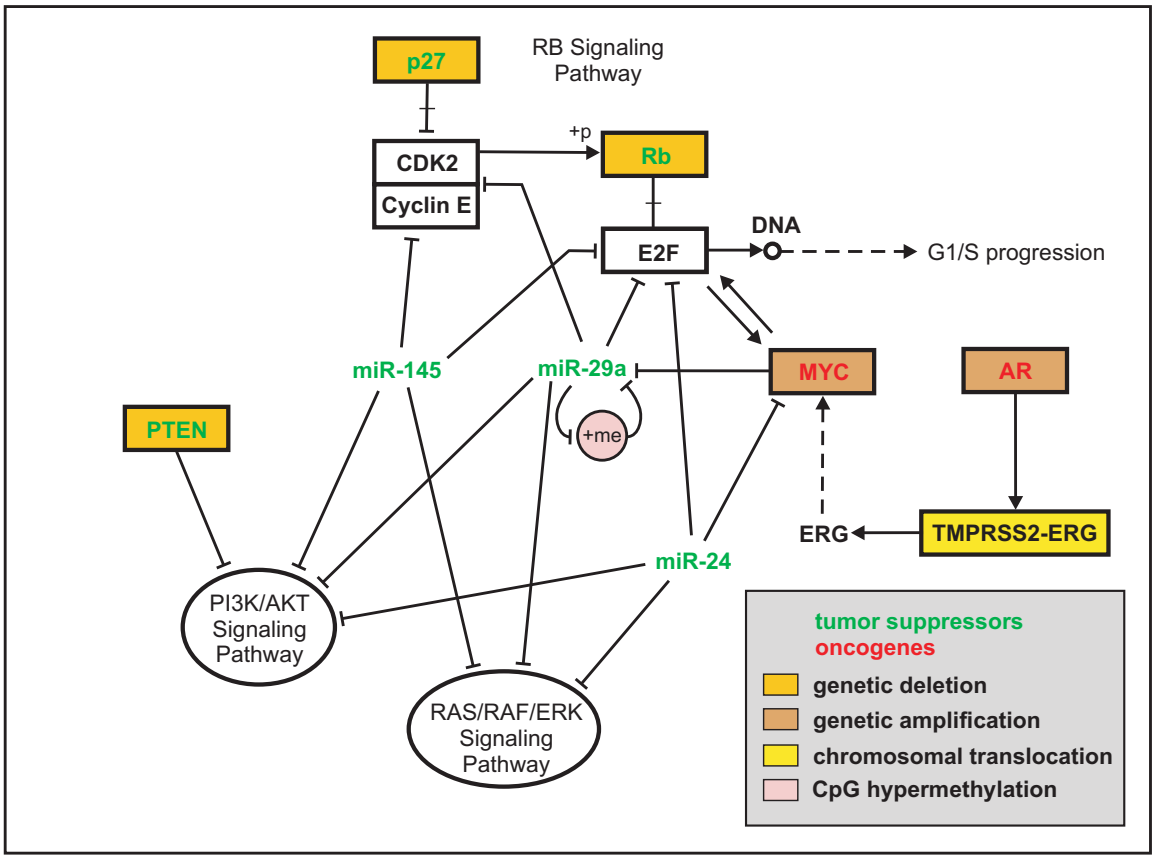

Figure 2.10: Integrative miRNA-mediated model for PCa development. A snapshot of a high-level integrative miRNA-mediated model for PCa development which encapsulates major sources of deregulation at the transcriptional, post-transcriptional, and signaling levels, coupled with genetic and epigenetic alterations.

TMPRSS2 [150]. Finally, based on the integration of CNA, transcriptome and mutation data, it was found that PI3K, RAS/RAF and RB1 signaling were commonly altered in primary tumors and metastases [150]. Moreover, it was stated that the data provided strong rationale for exploring the clinical activity of PI3K pathway inhibitors.

Interestingly, the findings depicted in Figs. 2.7 and 2.8 characterize miR-24, miR-29a, and miR-145, which are identified by IntegraMiR to be significantly downregulated, as inhibitors of the PI3K/AKT, RAS/RAF/ERK and RB1 signaling pathways through specific

FFLs and Type III loops, as depicted in Fig. 2.10, and suggest tumor suppressor roles for these miRNAs, coordinately cooperating with the tumor suppressors PTEN, CDKN1B (p27) and RB1 (Rb).

As a notable example, the Rb tumor suppressor gene product in Rb signaling is known to be a target of CDK2 (cyclin dependent kinase 2). When Rb is dephosphorylated, it interacts with E2F transcription factors and, in this way, prevents transcription of genes required for progression through the cell cycle. On the other hand, when Rb is phosphorylated by cell cycle dependent kinases, such as CDK2, it no longer interacts with E2F and the cell cycle proceeds through the G1-S checkpoint. The results depicted in Fig. 2.10 identify miR-29a and miR-145 as potential inhibitors of the CDK2/Cyclin E complex and E2F through FFLs and Type III regulatory loops and suggest that these miRNAs work in concert with p27 and Rb tumor suppressors, preventing passage from the G1 to the S phase.

In addition to the previously discussed genetic alterations and their effect on gene regulation, recent studies have found that miRNAs are both regulated epigenetically and play roles in epigenetic regulation of protein coding genes in different types of cancer, including PCa [71, 90, 96]. A recently validated example, which is relevant to our discussion, is miR-29a. It was discovered in [90] that the promoter region of miR-29a harbors numerous CpG sites. Moreover, it was determined that the experimentally measured methylation index of the miR-29a promoter was higher in the PCa cell group than in the prostate epithelial cell group, resulting in significant downregulation of miR-29a expression in PCa. More interestingly, miR-29a has been shown to play tumor suppressor roles by reciprocally

targeting DNA methyltransferases (DNMTs), which are key regulators of methylation of CpG islands [71, 96, 120].

We summarize these findings in the model depicted in Fig. 2.10, in which the red edges represent novel interactions predicted by IntegraMiR. In particular, edges emanating from the three miRNAs that target the two signaling pathways at the bottom represent the novel miRNA interactions depicted in Fig. 2.8. The resulting model suggests that upregulation of the oncogene *MYC* could take place due to genetic amplification and/or by *ERG* through *TMPRSS2-ERG* gene fusion. The upregulated MYC could then initiate a proliferative program, for instance, through the depicted MYC-E2F interaction, as well as by inhibiting the tumor suppressor miR-29a. In addition, other genetic and epigenetic alterations, for instance hypermethylation of miR-29a or deletion of *PTEN, p27 and Rb*, could further suppress the level of these tumor suppressor miRNAs and genes, leading to the activation of PI3K/AKT, RAS/RAF/ERK and RB signaling, and a consequent uncontrolled cellular growth.

It is important to emphasize at this point that miRNAs have attracted attention due to their diagnostic as well as therapeutic potential. Inactivating oncogenic miRNAs or restoring tumor suppressor miRNAs offers great prospects for cancer therapy [98, 99, 101, 107, 127]. As an important practical application, chromatin-modifying drugs, such as DNA methylation inhibitors, can be used to reactivate hypermethylated tumor suppressor miRNAs. Two DNMT inhibitors, 5-azacytidine and 5-aza-$2'$-deoxycytidine, have indeed been approved by the US Food and Drug Administration (FDA) for the treatment of myelodys-

plastic syndromes and acute myeloid leukemia [122].

2.4 Discussion and Conclusions

Earlier computational tools have focused primarily on identifying pairwise miRNA-target and TF-target interactions, either by relying on sequence-based analysis or expression data [8, 51, 126]. As a consequence, they may produce an excessively large number of false-positive predictions making them inefficient for experimental follow-up.

More recently, two promising methods have been proposed to identify miRNA/TF interactions [17, 174], which are based on the hypothesis that certain regulatory circuits, defined as motifs [3], appear in a statistically over-represented manner in the human and mouse genomes [155]. However, and for a given motif structure (e.g., an FFL), these methods attempt to predict all interactions (the three edges of an FFL) by utilizing a narrow set of computational tools and limited biological information. Although the methods can be employed to provide insights into the prevalence of various motif instances in gene regulatory networks, the user must keep in mind that the results may contain a rather large number of possibly unreliable predictions for experimental validation due to the fact that these methods do not effectively utilize certain known biological information to appropriately constrain and systematically reduce the resulting predictions.

In this research work, we introduced IntegraMiR, a novel computational method for inferring deregulated miRNA/TF-mediated regulatory loops and networks that appear in a

statistically over-represented manner in gene regulatory networks. IntegraMiR addresses the previous problems by appropriately constraining the statistical analysis of given mRNA/miRNA expression data and sequence-based target identification methods using relevant motif structures built by "prior" biological information readily available in existing databases. The main strength of IntegraMiR originates from its capacity to fuse information from multiple sources and incorporate several statistical techniques to exploit almost any accessible aspect of available information in the expression data to identify integrated regulatory loops and networks at the transcriptional, post-transcriptional and signaling levels. Therefore, IntegraMiR adds to the ongoing effort of developing effective computational techniques for network identification by utilizing available experimental data and existing biological knowledge in an effort to produce reliable predictions in a context-dependent manner.

To appropriately constrain the problem of predicting miRNA-target interactions, IntegraMiR focuses on specific types of three-node regulatory motifs and, in particular, FFLs that have attracted a great deal of attention in the literature. It is important to mention here that, in contrast to earlier work, such as that in [174], by identifying instances of deregulated FFL motifs and by using these motifs to construct interaction networks, IntegraMiR can also provide instances of two types of deregulated two-node motifs: miRNA-TF negative and double-negative feedback loops – see Figs. 2.6B, 2.7, and 2.9D.

IntegraMiR identified a number of already validated and novel deregulated miRNA/TF-mediated interactions. Although our interest was focused on certain types of regulatory

loops deregulated in PCa, the basic method can be easily modified to handle any other type of regulatory motif of interest and can be readily applied to other types of human disease, provided that appropriate miRNA and mRNA expression data are available. The results discussed in Section 2.3 demonstrate that IntegraMiR is a powerful computational tool for miRNA/TF-mediated network prediction, which can effectively result in novel hypotheses for further experimental study and validation. We should point out that the output results produced by IntegraMiR can be used by interested investigators to formulate additional hypotheses for experimental validation, beyond the ones discussed in this dissertation, which are expected to lead to additional novel findings.

IntegraMiR labels identified motifs into *consistent* and *inconsistent* loops, based on the rules depicted in Fig. 2.3 (see also Section 2.2.9). This is an additional piece of information that can be considered when evaluating the obtained results before carrying out experimental validation, when one seeks evidence based on expression data. As an illustrative example, we depict in Fig. 2.11 two loops considered in Section 2.3 – see Fig. 2.6A. These are instances of a Type I coherent FFL, with the green edges representing true-positive predictions and the red edge representing a novel interaction. The FFL depicted in Fig. 2.11A is identified by IntegraMiR to be consistently deregulated based on the data, whereas the FFL depicted in Fig. 2.11B is identified to be deregulated inconsistently.

The consistency of the deregulated FFL depicted in Fig. 2.11A implies that there is supporting evidence in the expression data to corroborate the intended reinforcing function modeled by this FFL. More specifically, when comparing tumor versus normal, the

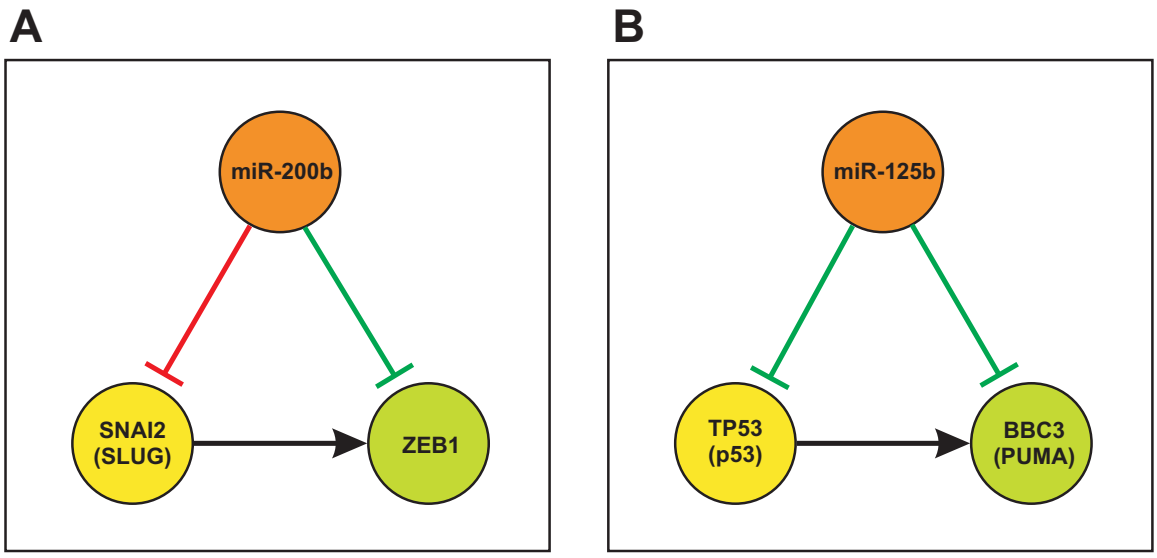

Figure 2.11: Examples of consistently and inconsistently deregulated FFLs identified by IntegraMiR. (A) A consistently deregulated Type I coherent FFL. (B) An inconsistently deregulated Type I coherent FFL. The green edges represent true-positive predictions whereas the red edge represents a novel prediction. The black edges represent known interactions.

observed significant upregulation of miR-200b leads to significant downregulation of the transcription factor SNAI2 (SLUG) and to a consequent downregulation of ZEB1. On the other hand, the inconsistency of the deregulated FFL depicted in Fig. 2.11B originates from the fact that, although the upstream inhibitor miR-125b is found by IntegraMiR to be significantly downregulated, and the opposite is true for the transcription factor TP53 (P53), the target gene *BBC3 (PUMA)* shows downregulation at the transcript level, which is contrary to the expected function modeled by this FFL.

Although all three interactions in an FFL, such as the one depicted in Fig. 2.11B, may have been experimentally validated individually, we may still not be able to observe consistent deregulation among the FFL nodes at the transcript level. This situation may occur due to a number of biological or technical factors. For example, the known miRNA-target

interactions available in miRTarBase may experimentally have been validated in certain cell type(s) and tissue(s) and may not take place in the context of interest (prostate tissue in our case). On the other hand, microarray experiments may not be able to capture the effect of translational repression by a miRNA (e.g., when this repression does not occur through mRNA degradation) or the fact that the mRNA level of a TF may not serve as a proxy for the corresponding protein-level activity. For example, in the case depicted in Fig. 2.11B, although miR-125b is downregulated and the transcription factor *TP53* transcript is upregulated based on the expression data, we may not have a high level of active TP53 protein in the nucleus that sufficiently correlates with the abundance of *TP53* mRNA transcripts. As a result, the target *BBC3* gene may not be transcribed in proportion to the level of the *TP53* transcript. In addition to the above, each node in an FFL may not necessarily participate only in that specific FFL and there can be numerous FFLs identified for certain nodes. This means that, by focusing on just one FFL, we may not be able to capture the relevant overall behavior. To do so, we may have to consider all collaborating FFLs in concert, which could potentially provide a more accurate and comprehensive representation of gene regulation for a specific gene of interest (we did this in several settings discussed in Section 2.3). Finally, alternate effects due to mechanisms other than FFL regulation, such as alterations at the genetic and epigenetic levels, could give rise to behaviors and observations that cannot be modeled by FFLs.

As we mentioned before, the two key hypotheses behind our interest in Type III loop motifs are that miRNAs play major roles in regulating signaling pathways due to their

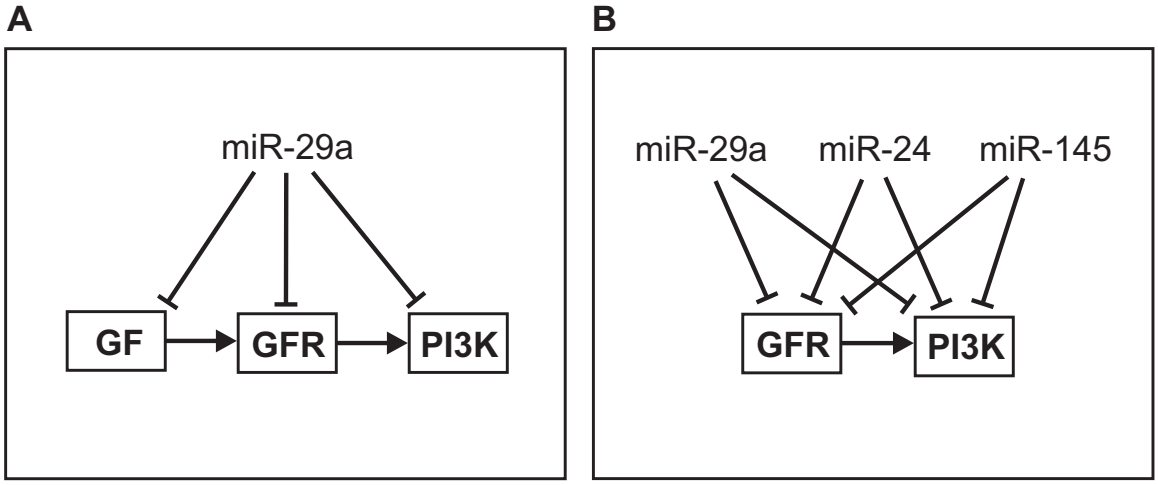

Figure 2.12: Complex regulatory motifs can be constructed from results obtained by IntegraMiR. (A) SIM motif of GF, GFR, and PI3K genes targeted by miR-29a in the KEGG prostate cancer pathway. (B) DOR motif of GFR and PI3K co-targeting by miR-29a, miR-24, and miR-145 in the KEGG prostate cancer pathway.

sharp dose-sensitive nature, and that targets of single miRNAs are more connected (i.e., interact) at the protein level than expected by chance. IntegraMiR identifies closely related miRNA targets on pathways deemed to be important in PCa and delineates certain miRNA-mediated three-node regulatory loops in the KEGG Prostate Cancer Pathway. As an example, we refer to the two consecutive Type III loops for miR-29a depicted in Fig. 2.12A, which have been constructed from the results depicted in Fig. 2.8. The obtained mechanism of a single miRNA regulating several closely related genes typically working together to perform a common task represents a single-input module (SIM) motif [3]. SIMs can partially explain how individual miRNAs could be potent regulators of pathway activity even though the effect of the miRNA on any single gene target may be modest [6, 131, 156].

It has also been demonstrated in [156] that targeting of a set of genes by multiple miRNAs could produce effects that are much more dramatic than the modest effects exerted by

individual miRNAs. A notable example identified by IntegraMiR in the KEGG Prostate Cancer Pathway is the co-targeting of *GFR* and *PI3K* genes by miR-29a, miR-24 and miR-145 depicted in Fig. 2.12B (which has been constructed from the results depicted in Fig. 2.8). The resulting network structure represents a dense overlapping regulon (DOR) motif [3] in which several input miRNAs co-regulate a set of output genes (known as a regulon). Co-targeting in a DOR pattern presumably strengthens the notion that the miRNAs involved share similar regulatory roles. It is noteworthy that IntegraMiR can identify numerous examples of miRNA co-targeting in the context of FFLs as well – see Fig. 2.7. Clearly, the three-node loop motifs considered in this dissertation can serve as basic building blocks for identifying more complex regulatory motifs, such as SIMs and DORs [2, 26].

In principle, discoveries obtained by integrative computational approaches, similar to IntegraMiR, can provide systemic insights into the molecular biology of miRNA-mediated interactions and can, thereby, assign context-dependent biological functions to poorly understood roles of miRNAs.

Chapter 3

MicroRNA-mediated Networks in Autism Spectrum Disorders

Introduction

In Chapter 2, we looked into how certain gene regulatory loops can provide systemic insights into the molecular biology of miRNA-mediated interactions and can, thereby, assign context-dependent biological functions to poorly understood roles of miRNAs. Consistent with the recent findings on miRNA networks, we argued that effective drug targeting and successful disease treatments will eventually be realized through these molecular mechanisms underlying physiological and pathological conditions of interest and that miRNAs pose promising potential in this context.

CHAPTER 3. MICRORNA-MEDIATED NETWORKS IN AUTISM SPECTRUM DISORDERS

In this chapter, we will investigate, as an application of our previous work on miRNA/TF loop and network identification, the role of LIN28-regulated miRNAs and their networks in Autism Spectrum Disorders (ASDs). Recent discoveries have provided experimental evidence that miRNAs could also play a crucial role in the post-transcriptional regulation of gene expression in neurons and diseases associated with neuronal growth and synaptic function. These include autism spectrum disorders, as the category of our interest.

To this end, we will first provide the biological background of this research. Then, we will propose a novel biological hypothesis whose validity we will investigate by combining our computational research work together with the experimental work of our collaborators (Dr. Mollie Meffert's lab at the School of Medicine). Next, we will present a bioinformatics analysis we carried out in order to assess the statistical significance of enrichment of predicted targets of LIN28-regulated miRNAs in ASD-related genes. We performed this analysis using publicly available datasets, which include gene expression levels in distinct contexts of interest as well as computationally predicted miRNA targets. Our objective was to provide evidence for guiding costly and time-consuming experiments for validation. We also used experimental data, provided by our collaborators, to validate part of our computational analysis. At the end of this chapter, we will discuss the clinical utility and translational impact of this research and provide further discussion and conclusions.

3.1 Biological Background

It is well known that the regulation of gene expression at the level of translation is a critical factor in the neuronal response to several stimuli, including synaptic activity [60] and neurotrophins [130], among others. A well-studied example of such a stimulus is the Brain-Derived Neurotrophic Factor (*BDNF*), which is broadly expressed in the mammalian brain and critically contributes to modifications of synaptic growth and function. BDNF selectively targets an estimated 4% or less of expressed mRNAs that undergo enhanced translation [130], which we discuss in more detail later in this section. Moreover, it was experimentally determined in [59] that the function of the miRNA biogenesis pathway plays a pivotal role in BDNF-mediated regulation of translation. Here, we will briefly review an experimentally identified and validated mechanism that mediates target specificity in BDNF-regulated translation. This review will lay the biological foundation of our research work.

To investigate the role of miRNAs in BDNF translation specificity, it was examined in [59] whether BDNF itself might affect the miRNA biogenesis pathway. By using miRNA arrays, an overall pattern was observed toward higher miRNA abundance in BDNF-treated versus mock-treated primary neurons, which indicates that BDNF might regulate a key component of miRNA biogenesis, such as the DICER processing complex. It was indeed validated in [59] that BDNF causes a distinct transcription-independent increase in DICER levels in BDNF-stimulated neurons. As a matter of fact, it is known that BDNF binds to TRKB receptors and this binding triggers growth- and survival-promoting pathways,

including the PI3K/AKT and MAPK/ERK signaling pathways. A previous research in tumor cell lines reported that a component of the DICER complex, HIV-1 TAR RNA-binding protein (TRBP), could undergo ERK-dependent phosphorylation and this could lead to the stabilization and enhancement of DICER levels [113]. It was also verified in [59] that BDNF rapidly causes the induction of phospho-ERK and a multi-banding pattern of TRBP. This can explain the observed elevated levels of DICER, which could invoke mature miRNA biogenesis. With elevated levels of miRNAs, additional mRNAs could be targeted for repression.

It was also examined in [59] whether the miRNA biogenesis pathway could also be regulated to positively select BDNF-upregulated targets in protein synthesis. Based on miRNA array experiments, a small number of miRNAs were found to be decreased in response to BDNF, including several members of the let-7 family of miRNAs.

In addition, it is known that miRNA biogenesis can be regulated at several steps by trans-acting factors, such as the LIN28 RNA-binding proteins [55], which target let-7 family members. Once LIN28 binds to the pre-miRNA molecule, it causes the uridylation of the molecule. As a result, it suppresses processing of targeted pre-miRNA to mature miRNA [55]. This could in fact present a mechanism to reduce specific mature miRNAs even when the DICER level is elevated. Notably, it was experimentally observed in [59] that, following BDNF exposure, a rapid and transcription-independent increase in LIN28a takes place.

Previous research showed that the terminal loop of each of let-7, miR-107 and miR-143

pre-miRNAs has a functionally confirmed "GGAG" sequence motif that is recognized by LIN28 [50, 55]. By using individual qRT-PCR assays, it was observed that BDNF induces a significant and reproducible decrease in the abundance of all tested members of the let-7 family, as well as miR-107 and miR-143, even though not all decreases were reproducibly detected using less sensitive miRNA assays.

The previous results led to the following hypothesis: if LIN28 positively selects BDNF-upregulated targets by binding to specific pre-miRNAs and decreasing their mature miRNAs, then a mRNA transcript that contains functional sites for a LIN28-downregulated miRNA would be expected to show BDNF-enhanced translation. To test this hypothesis, the existence of binding sites for LIN28-regulated miRNAs in the 3'UTR regions of mRNA transcripts was examined in [59], known to undergo upregulation, downregulation or no change in terms of the level of their translation. It was found that thirteen representative BDNF-upregulated targets contain two or more binding sites for a LIN28-regulated miRNA. However, BDNF-downregulated or unregulated targets did not contain such sites [59]. To directly test the role of LIN28a in BDNF target mRNA selection, it was experimentally validated in [59] that depletion of LIN28a, through RNAi or LIN28a knockdown, prevented the enhanced translation of representative mRNA targets that are normally upregulated by BDNF. Moreover, these targets, which are normally derepressed and upregulated by BDNF, were observed to remain repressed in LIN28a-deficient neurons.

As a consequence of these results, a coordinated mechanism was established for genome-wide control of translation specificity that involves BDNF-dependent positive and nega-

tive regulation of the miRNA biogenesis pathway, which involves the combined action of BDNF on DICER and LIN28a [59]. Although it is possible that alternative mechanisms may coexist, the experimental results obtained in [59] strongly suggest that the dual control of the miRNA biogenesis pathway by BDNF through LIN28a and DICER plays a crucial role in selectively determining both upregulated and downregulated targets in BDNF-mediated translation.

3.2 Biological Hypothesis

In [59], it was demonstrated that LIN28 was required for neurotrophin-induced translation of suites of genes that collectively contribute to synaptic growth and plasticity. In particular, LIN28 selectively binds to and causes the degradation of the let-7 family of miRNAs that suppress many pro-growth genes under basal conditions. This function of one of the let-7 family members, i.e. let-7a, is consistently demonstrated in Fig. 2.7 of Chapter 2 in the context of prostate cancer. It was discussed in that chapter that the let-7a miRNA, being present in certain coherent Feed-Forward Loops, suppresses *MYC-E2F1* genes that promote cell growth. As a result, our previous research work identified let-7a as an anti-proliferative miRNA.

On the other hand, it was experimentally validated in [59] that the prevention of LIN28-mediated downregulation of let-7 family of miRNAs can completely block neuronal growth responses to BDNF. Congruent with this observation, the LIN28/let-7 axis has been recog-

nized as a highly evolutionarily-conserved pathway that controls growth and development. Interestingly, recent experimental investigation by our collaborators has revealed dysregulation of the LIN28/let-7 axis in a mouse model of Fragile X Syndrome (FXS). FXS is an autism spectrum disorder disease, and as such, it is the most commonly inherited form of mental disability. It is caused by the expansion of CGG repeats in the promoter of *FMR1*, which leads to the hypermethylation and transcriptional silencing of the *FMR1* gene product, the Fragile X mental retardation protein (FMRP). FMRP is essential for normal cognitive development since it binds to a large number of mRNAs in neurons and regulates targets in the hub of key signaling pathways.

The *FMR1* knockout (KO) mouse model for FXS recapitulates many of the synaptic and behavioral phenotypes found in FXS, including synaptic/neuronal overgrowth and cognitive disorders. Using this model, and based on the previous biological facts and findings, we plan in this research work to investigate the following novel biological hypothesis:

> Dysregulation of LIN28-regulated miRNAs may lead, through their network of interactions, to a selective overabundance of growth-promoting synaptic proteins that could account for synaptic and cognitive functions in FXS.

3.3 Methods

To investigate the previous hypothesis, we focused on answering the following questions:

(i) Are predicted targets of LIN28-regulated miRNAs enriched in autism-related genes?

(ii) Can we identify type III loops that comprise LIN28-regulated miRNAs and genes that are directly related to and significantly upregulated in the disease state? Can we construct a network of interactions of these miRNAs and genes? Among the predicted targets of LIN28-regulated miRNAs, which directly related and upregulated genes will be in the predicted network? What is the structure of this network in terms of coordinated functions of LIN28-regulated miRNAs?

(iii) Can the predicted LIN28-regulated miRNA-target interactions in the computationally constructed networks be experimentally validated? Validation requires a series of focused experiments that would provide supporting experimental evidence as to whether the dysregulation of LIN28-regulated miRNAs is indeed contributing to the dysregulation of growth-promoting synaptic proteins, and therefore contribute to the disease state.

To answer the question in (i), we need to assess the statistical significance of enrichment of predicted targets of LIN28-regulated miRNAs in ASD-related genes. To accomplish this task, we perform two distinct sets of analyses based on independent studies and databases in the context of ASDs. For this purpose, we use the R software package WGCNA to perform the enrichment analysis based on the hypergeometric test [82].

The conceptual reasoning behind using the hypergeometric test is that, there is some reference set of genes, and that these genes can be divided into two classes: those that are interesting (autism-related for example), and those that are not. In addition, there are

other important gene characteristics, such as a gene belonging to a particular category, for example by being a predicted target of a given miRNA. Moreover, one would like to ask whether there is an association between a gene being interesting and having a particular property. In this case, the hypergeometric test quantifies the significance of enrichment of genes having that particular property among those interesting genes by a P-value [82]. We perform this test for gene sets representing available predicted targets of the let-7 family of miRNAs (the 9 miRNAs let7a-g, let7i, as well as miR-98). We also perform this test for the gene set representing predicted targets of miR-9, and we do the same for miR-107, as well as for miR-143, three additional miRNAs known to be regulated by LIN28. In addition, we perform the test for the gene set obtained by combining all genes predicted to be targeted by the previous 12 LIN28-regulated miRNAs of interest. Finally, we perform the test for miR-122, a liver-specific miRNA that is not regulated by LIN28, which we use as a control.

The reason we also use the combined set of genes is because, although miRNAs are known to have subtle effects on protein levels of individual targets, given the multiplicity of their targets and the concurrent downregulation of several of these targets, their cumulative influence can significantly affect the outcomes controlled by cellular pathways. We looked into several examples of this kind in the context of regulatory loops in Chapter 2 - see Figs. 2.6-2.12. With this known fact, we expect to find that the predicted targets of a specific miRNA may not be enriched in a particular set of genes related to a certain cellular function. However, when we incorporate the targets of other miRNAs cooperating with this miRNA for the same cellular function, we may find that the collective set is enriched

in the gene set of interest, e.g. autism-related genes based on a specific study or database.

For each of the two autism-related enrichment analyses we perform here, we employ three types of gene lists: i) a reference list of expressed genes, ii) lists comprised of genes that are predicted to be miRNA targets as we discussed above, and iii) a test list of genes that are associated with autism. Each analysis will focus on discovering enrichment of a target gene set in the test gene list corresponding to the particular analysis.

We considered two reference gene lists that we deemed to be reliable and relevant to the study. The reference list we used in our first analysis was obtained from a recent RNASeq expression study at Johns Hopkins [49], which comprises 57 control and 47 autism samples. By following a practical approach to exclude genes that show negligible expression levels, we included only those genes in the reference gene list that exhibit a normalized and transformed read count value greater than 1 in at least 10% of the available samples. This resulted in a reference list comprising 17,693 expressed genes, which we refer to as REFLIST1.

On the other hand, we obtained the reference gene list we used in our second analysis from the Brainspan database (www.brainspan.org), which comprises 525 samples associated with the hippocampus or the cortex of young and old individuals. We included in this list only genes that exhibit an RPKM value greater than 1 in at least 10% of the samples. This resulted in a list comprising 14,890 expressed genes, which we refer to as REFLIST2.

To construct the predicated gene lists, we extracted the sequence-based predicted targets of the 12 LIN28-regulated miRNAs of interest from the TargetScan database (v.7)

(http://www.targetscan.org), and selected genes based on their 3UTR region and a seed length of 7.

Finally, we constructed two test gene lists. For the first enrichment analysis, we considered a list of 749 differentially expressed genes in autism, obtained from [49]. As required by the hypergeometric test, we matched these genes with the genes in REFLIST1, which was also obtained from the same study. By doing so, we obtained 702 differentially expressed genes that formed our first test list, which we refer to as the HOPKINS list. For the second analysis, we considered a list of 631 autism-associated genes, obtained from the Simons Foundation Autism Research Initiative (SFARI) database (SFARI.org). We then matched these genes with the genes in REFLIST2 and obtained 550 autism associated genes that formed our second test list, which we refer to as the SFARI list.

Before proceeding with our first enrichment analysis, we visualize in Fig. 3.1 the overall enrichment of predicted targets of the 12 LIN28-regulated miRNAs we consider in this study versus non-targets in autism-related genes. Moreover, we depict in Fig. 3.2, the number of autism-related genes in the HOPKINS list that are predicted to be targeted by a miRNA of interest. These results indicate that there is some enrichment of predicted targets of LIN28-regulated miRNAs in autism-related genes, although it is not clear whether this enrichment is statistically significant. However, by performing statistical analysis using the hypergeometric test, and by including miR-122 as control, we obtain the results depicted in Table 3.1. All target sets, except the ones corresponding to let-7 and miR-122, show statistically significant enrichment in autism-related genes (Bonferroni adjusted p-value

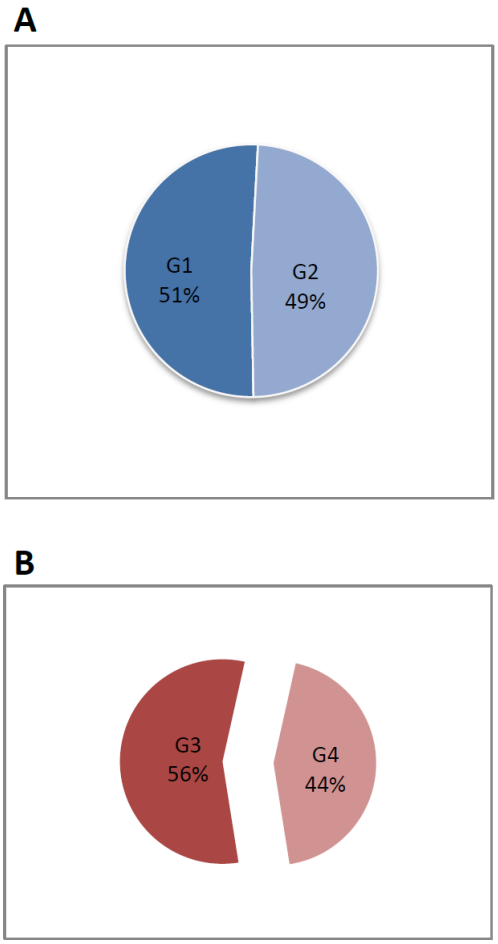

Figure 3.1: (A) Proportion of genes predicted to be targeted by the LIN28-regulated miRNAs in the REFLIST1 gene list. G1: Predicted targets of LIN28-regulated miRNAs of interest in REFLIST1. G2: Genes in REFLIST1 that are not predicted to be targeted by the LIN28-regulated miRNAs. (B) Proportion of autism-related genes predicted to be targeted by the LIN28-regulated miRNAs in the HOPKINS gene list. G3: Autism-related genes in the HOPKINS list that are predicted to be targeted by the LIN28-regulated miRNAs. G4: Autism-related genes in the HOPKINS list that are not predicted to be targeted by the LIN28-regulated miRNAs. There are 17,693 genes in REFLIST1 and 702 genes in the HOPKINS list.

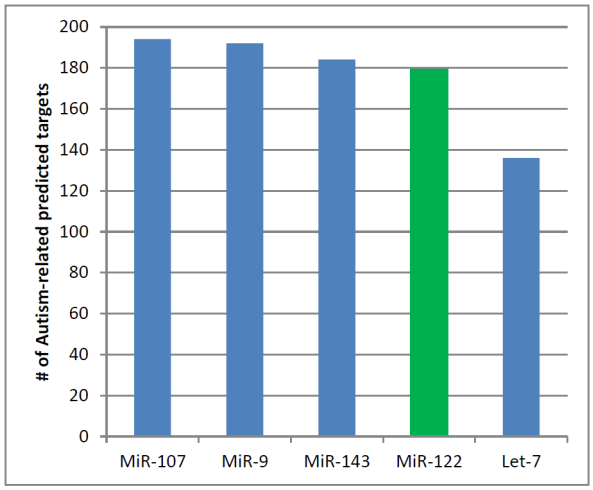

Figure 3.2: Number of autism-related genes (in descending order) in the HOPKINS list predicted to be targeted by the miRNAs of interest. MiR-122 is used as a control in the enrichment analysis.

< 0.05). Notably, Fig. 3.2 indicates that miR-122 is predicted to target a larger number of autism-related genes than let-7. However, the significance of enrichment of its target set is lower than that of let-7 according to Table 3.1. This reinforces our expectation that the target set of a given miRNA that targets a large number of autism-related genes is not necessarily enriched in these genes.

It is important to note here that, although the significance level of enrichment of the predicted targets associated with let-7 is slightly above 0.05, there are 9 miRNAs in this family. As we discussed in Chapter 2, miRNAs co-targeting a group of genes are expected to have an appreciably larger effect on regulating their target set through their cumulative influence, which is not captured by the hypergeometric test.

Table 3.1: Statistical significance of enrichment of predicted targets of LIN28-regulated miRNAs in autism-related genes based on the HOPKINS list, as determined by the hypergeometric test. P-values have been adjusted for multiple testing using Bonferroni correction. The "pooled" gene set is formed by combining the predicted targets of all 12 LIN28-regulated miRNAs we consider in this study, whereas Let-7 indicates the gene set of predicted targets of all 9 let-7 family miRNAs (including miR-98). The target set associated with miR-122 was used as control.

MiRNA Target Set	Number of Overlaps	P-value	Adjusted P-value
Pooled	393	6.88E-06	4.13E-05
MiR-9	192	8.76E-05	5.26E-04
MiR-107	194	1.10E-04	6.58E-04
MiR-143	184	7.28E-03	4.37E-02
Let-7	136	9.57E-03	5.74E-02
MiR-122	179	1.35E-02	8.07E-02

Since the target gene sets corresponding to LIN28-regulated miRNAs have been found to be significant, the previous results provide supporting computational evidence that LIN28-regulated miRNAs could potentially be involved in the regulation of autism-related genes. As a consequence, dysregulation of these miRNAs and their network of interactions could be associated with this disease.

To seek further evidence for the role of LIN28-regulated miRNAs in autism, we performed a second analysis based on publicly available gene sets. In particular, we investigated enrichment of predicted targets of LIN28-regulated miRNAs in autism-related genes using the Brainspan and SFARI databases. Notably, REFLIST2 includes 1,226 genes that are not included in REFLIST1, a significant number of genes which is more than double the number of genes in either one of the test lists. More importantly, however, 596 out of

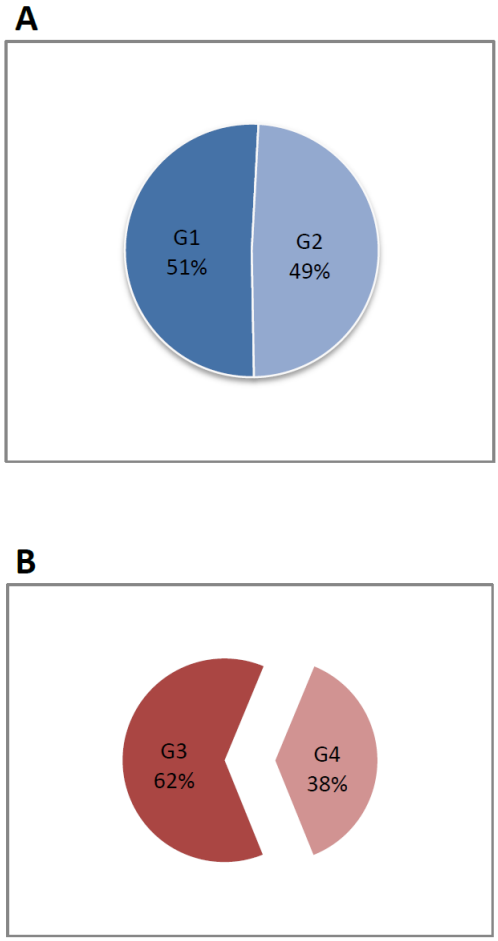

Figure 3.3: (A) Proportion of genes predicted to be targeted by the LIN28-regulated miRNAs in the REFLIST2 gene list. G1: Predicted targets of LIN28-regulated miRNAs of interest in REFLIST2. G2: Genes in REFLIST2 that are not predicted to be targeted by the LIN28-regulated miRNAs. (B) Proportion of autism-related genes predicted to be targeted by the LIN28-regulated miRNAs in the SFARI gene list. G3: Autism-related genes in the SFARI list that are predicted to be targeted by the LIN28-regulated miRNAs. G4: Autism-related genes in the SFARI list that are not predicted to be targeted by the LIN28-regulated miRNAs. There are 14,890 genes in REFLIST2 and 550 genes in the SFARI list.

the 631 autism-related genes obtained from SFARI are not among the 749 differentially-expressed genes in autism obtained from [49], with the two sets having only 35 common genes.

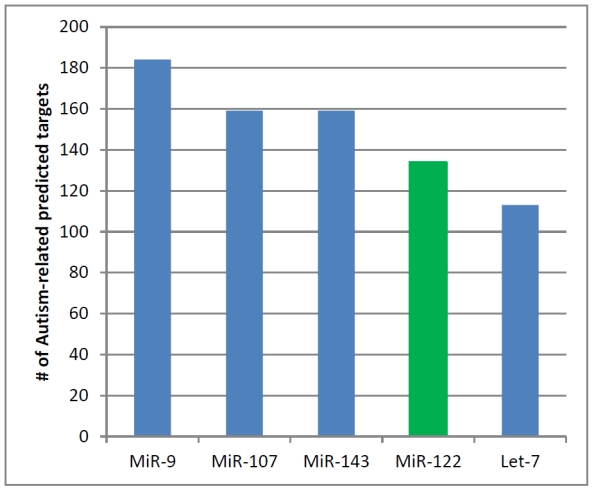

Figure 3.4: Number of autism-related genes (in descending order) in the SFARI list predicted to be targeted by the miRNAs of interest. MiR-122 is used as a control in the enrichment analysis.

Similar to what we did in the first analysis, we visualize in Fig. 3.3 and Fig. 3.4 the overall enrichment of predicted targets of LIN28-regulated miRNAs versus non-targets in autism-related genes using REFLIST2 and the SFARI list. The results are consistent with the ones obtained using REFLIST1 and the HOPKINS list, suggesting again that there is some enrichment of predicted targets of LIN28-regulated miRNAs in autism-related genes. In terms of statistical significance, the results depicted in Table 3.2 show once more that the gene sets corresponding to LIN28-regulated miRNAs are significantly enriched in autism-related genes, although this now is not true for the target gene set corresponding to let-7 when the p-values are adjusted for multiple testing using Bonferroni correction. These additional results are particularly encouraging, given the small overlap between the SFARI and HOPKINS datasets, and provides further supporting evidence of our computational findings.

Table 3.2: Statistical significance of enrichment of predicted targets of LIN28-regulated miRNAs in autism-related genes based on the SFARI list, as determined by the hypergeometric test. P-values have been adjusted for multiple testing using Bonferroni correction. The "pooled" gene set is formed by combining the predicted targets of all 12 LIN28-regulated miRNAs we consider in this study, whereas Let-7 indicates the gene set of predicted targets of all 9 let-7 family miRNAs (including miR-98). The target set associated with miR-122 was used as control.

MiRNA Target Set	Number of Overlaps	P-value	Adjusted P-value
MiR-9	184	4.38E-09	2.63E-08
Pooled	343	4.15E-08	2.49E-07
MiR-107	159	1.31E-03	7.88E-03
MiR-143	159	4.23E-03	2.54E-02
Let-7	113	2.62E-02	1.57E-01
MiR-122	134	3.78E-01	1.00E+00

3.4 Results

3.4.1 Predicted LIN28-regulated MiRNA-target Interaction Networks in FXS

Given the results we obtained from our previous bioninformatics analysis, we sought to examine whether we could identify predicted targets of LIN28-regulated miRNAs among relevant genes reported in the literature to be upregulated in the disease state. By doing so, we could provide answers to the second group of questions in Section 3.3.

We have mentioned before that, for the mouse model of FXS, the FMR1 knockout (KO)

mouse serves as an appropriate and desirable animal model since it recapitulates many of the synaptic and behavioral phenotypes found in FXS, including synaptic/neuronal overgrowth and cognitive disorders. Interestingly, in a very recent study with regards to FXS, systematic protein expression measurements in neocortical synaptic fractions from FMR1 KO and wild-type (WT) mice have been reported [149]. These measurements revealed upregulated proteins, which are associated with previously unidentified and known genes involved in synapse formation, function, and brain development, as well as other genes linked to mental disability and autism.

Therefore, and with our objective to investigate the effect of dysregulation of LIN28-regulated miRNAs in the overabundance of genes involved in synapse formation and function, we decided to identify the predicted targets of LIN28-regulated miRNAs among certain highly relevant categories of genes. Specifically, we considered 6 representative protein complexes that have been identified in [149] to play key roles in pre- and post-synaptic organization of signaling complexes, as well as in determining the structure and function of nervous system development. These protein complexes were obtained by mapping the genes associated with the dysregulated proteins to the CORUM database [124], which contains curated data on 453 mouse protein complexes.

Table 3.3: List of 17 genes whose products form 6 representative protein complexes that have been identified to play key roles in pre- and post-synaptic organization of signaling complexes, as well as in determining the structure and function of nervous system development. Genes highlighted in red have been identified in this Dissertation to participate in the LIN28-regulated miRNA-target interaction network depicted in Fig. 3.5.

ACTN1	DNM1	**SHANK2**	SYNGAP1
CAMK2A	**GRIN2B**	**SNAP25**	VAMP2
CPLX2	**HOMER1**	**STX1A**	
DLG1	**PSD3**	SYN1	
DLG4	**SHANK1**	SYN2	

Using the previous information, we identified the predicted targets of LIN28-regulated miRNAs among the 17 genes listed in Table 3.3, whose products form the previously discussed complexes and whose protein levels were found to be significantly upregulated in FMR1 KO versus WT samples. Based on our results, we constructed a predicted miRNA-target interaction network that parallels the logic behind the Type-III loop construction in Chapter 2, which we depict in Fig. 3.5. Notably, 9 out of 17 significantly upregulated targets were found to be putative targets of at least one LIN28-regulated miRNA.

In addition to the above, 19 genes (see Table 3.4) were reported in [149] whose products: i) were found to be upregulated in KO versus WT samples, ii) existed in the gene-to-cognition postsynaptic proteome (G2Cdb:PSP) database (http://www.genes2cognition.org) listing the core set of synaptic proteins, iii) were among the 842 direct FMRP mRNA targets identified by cross-linking immunoprecipitation and RAN-seq analyses in [27], iv) were included among the autism-associated genes in the SFARI database. Note that FMRP is the protein product of *FMR1* gene, the silencing of which leads to the development of FXS.

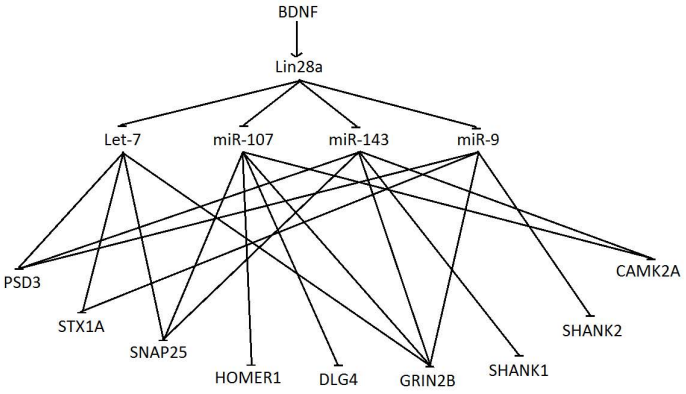

Figure 3.5: Predicted LIN28-regulated miRNA-target interactions using 17 selected up-regulated genes, whose products form 6 representative protein complexes that play key roles in the nervous system development and whose protein levels were found to be significantly upregulated in FMR1 KO versus WT samples. Let-7 represents the let-7 family of 9 miRNAs (including miR-98).

Intriguingly, we identified 14 out of these 19 genes to be predicted targets for at least one of the LIN28-regulated miRNAs of interest and constructed the corresponding predicted miRNA-target interaction network for these genes, which we depict in Fig. 3.6.

Together, these results provide strong evidence for our biological hypothesis discussed in Section 3.2 (and the subsequent experimental work) that dysregulation of LIN28-regulated miRNAs may lead to a selective overabundance of growth-promoting synaptic proteins that could account for synaptic and cognitive functions in FXS.

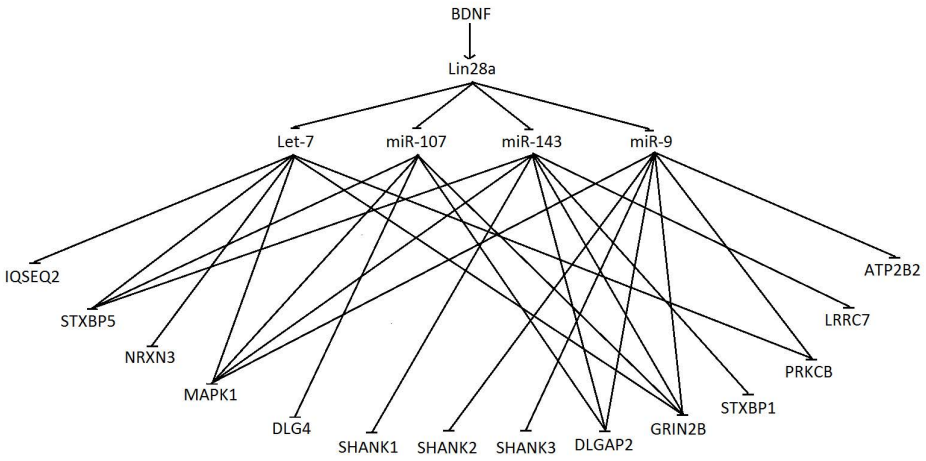

Figure 3.6: Predicted LIN28-regulated miRNA-target interactions using 19 core synaptic genes in FXS associated with autism, which are known to be bound at the mRNA level by FMRP and upregulated in FMR1 KO samples. Let-7 represents the let-7 family of miRNAs (including miR-98).

Table 3.4: List of 19 genes whose products were found to be upregulated in KO versus WT samples, existed in the gene-to-cognition postsynaptic proteome (G2Cdb:PSP) database listing the core set of synaptic proteins, were among the 842 direct FMRP mRNA targets identified by cross-linking immunoprecipitation and RAN-seq analysis, and were included among the autism-associated genes in the SFARI database. Genes highlighted in red have been identified in this Dissertation to participate in the LIN28-regulated miRNA-target interaction network depicted in Fig. 3.6.

ATP2B2	KCNMA1	**PRKCB**	**STXBP5**
DLG4	**LRRC7**	**SHANK1**	SYN1
DLGAP2	**MAPK1**	**SHANK2**	SYNE1
GRIN2B	**NRXN3**	**SHANK3**	SYNGAP1
IQSEC2	PRICKLE2	**STXBP1**	

3.4.2 Supporting Experimental Findings

Our collaborators have been working on utilizing the previous predicted miRNA-mediated interaction networks with a goal to answer the third group of questions in Section 3.3 and develop a biomarker for autism, as well as possible therapeutic strategies for this disease. One of their initial steps was to investigate the expression levels of a subgroup of the miRNAs of interest (i.e., let-7a, let-7f, and miR-9), as well as the protein levels of LIN28a, TRBP and DICER, which play crucial roles in the dual mechanism of regulation induced by BDNF (see Fig. 1.1).

To this end, we constructed a predicted network of interactions among the LIN28-regulated miRNAs, LIN28a, TRBP, and DICER, which we depict in Fig. 3.7. As we discussed in the Introduction of this chapter, it is known that BDNF positively regulates a group of genes through the LIN28a axis, which represses the 12 miRNAs of interest, and negatively regulates other genes by inducing TRBP and DICER. This network, however, suggests that the 12 LIN28-regulated miRNAs mutually inhibit the dual regulation mechanism induced by BDNF (see Fig. 1.1).

For the three miRNAs let-7a, let-7f, and miR-9, our collaborators obtained experimental measurements through qRT-PCR experiments. They also measured the protein levels of the three genes *LIN28a, TRBP,* and *DICER* by western blot. For these six molecular species, we used the Wilcoxon test to assess the statistical significance of the level of their dysregulation in the disease state (*FMRP* knockdown) versus control, with the results depicted in Table 3.5.

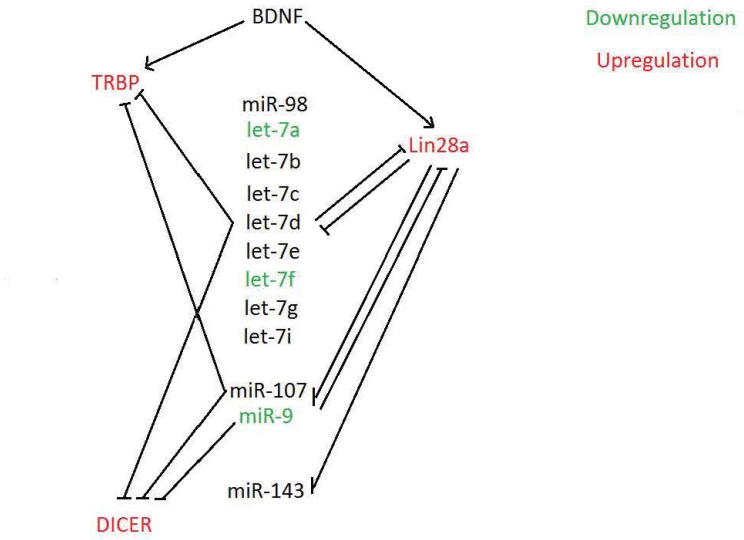

Figure 3.7: Predicted LIN28-regulated miRNA-target interactions in the context of Type-III loops regulating DICER, TRBP and LIN28a.

It turns out that the experimental data obtained through *FMRP* knockdown experiments are consistent with the predicted interactions for which data is available, see Fig. 3.7. Specifically, an upregulation in LIN28a, TRBP, and DICER is consistent with the down-regulation in the levels of miRNAs regulated by LIN28. These interactions consist of 7 Type-III loops with overlapping edges, listed in Table 3.6. To calculate a significance score for each Type-III loop, we used Fisher's method in the same way as we did in Chapter 2. In this way, we obtain a systemic view on how these molecular species interact in the context of the predicted network and this yields a holistic understanding of the dysregulated network.

Table 3.5: Dysregulation of molecular species in the induced disease state (*FMRP* knockdown) versus control. Expression levels for miRNAs are obtained by qPCR and protein levels are measured by western blot.

Molecular Species	Deregulation	Number of Samples (Control)	Number of Samples (Disease State)	P-value (Wilcoxon Test)
Let-7a	Downregulated	3	3	5.00E-02
Let-7f	Downregulated	3	3	5.00E-02
MiR-9	Downregulated	3	3	5.00E-02
LIN28a	Upregulated	6	9	8.00E-04
DICER	Upregulated	2	4	6.67E-02
TRBP	Upregulated	5	6	8.66E-03

Table 3.6: Predicted dysregulated Type III loops in the induced disease state (*FMRP* knockdown) versus control. The significance score has been calculated in the same way as in Chapter 2.

MiRNA	Gene 1	Gene 2	Type III Loop Significance Score
Let-7a	TRBP	LIN28a	4.38E-05
Let-7f	LIN28a	TRBP	4.38E-05
Let-7a	DICER	LIN28a	2.57E-04
Let-7f	LIN28a	DICER	2.57E-04
MiR-9	DICER	LIN28a	2.57E-04
Let-7f	TRBP	DICER	1.91E-03
Let-7a	DICER	TRBP	1.91E-03

3.5 Discussion and Conclusions

In this collaborative research work, our goal was to investigate the biological hypothesis stated in Section 3.2 and address the three groups of questions listed in Section 3.3.

For the purpose of enrichment analysis, we first considered two independent reference gene sets that were relevant to the study and were deemed to be reliable. The first set was obtained from a recent RNA-Seq study at Johns Hopkins [49], whereas the second reference set was obtained from the Brainspan database, as we discussed in Section 3.3. We then calculated the statistical significance of enrichment of the predicted targets of miRNAs of interest using two slightly overlapping test gene lists: a list obtained from [49], comprising genes that are differentially expressed in autism, and a list obtained from the SFARI database, comprising autism-related genes.

Both of our analyses consistently demonstrated that the predicted targets of LIN28-regulated miRNAs of interest collectively have statistically significant enrichment in autism-related genes.

An interesting observation of our work is that the relative significance levels of enrichment of predicted targets of miR-9, miR-107, and miR-143 is consistent between the two analyses (see Tables 3.1 and 3.2) despite the fact that the test gene sets used in these two analyses have small overlap. However, there is a difference between the relative numbers of autism-related predicted targets of miR-9 and miR-107 (see Figs. 3.2 and 3.4).

In terms of the predicted interaction networks for LIN28-regulated miRNAs, we should note that the experimental results are consistent with our computational results based on

enrichment analyses. Altogether, these results provide a strong evidence for our biological hypothesis. Moreover, subsequent experimental work may provide further evidence that the dysregulation of LIN28-regulated miRNAs may lead to a selective overabundance of growth-promoting synaptic proteins that could account for synaptic and cognitive functions in ASDs.

In terms of translational impact, clinical relevance, and significance, it is noteworthy that Fragile X Syndrome (FXS) represents the most common inherited form of mental disability affecting 1 in 2500 births. FXS is also the leading inherited cause of autism and other mental disabilities, whereas, diagnosis for these disabilities is made by clinical evaluation. Reliance upon clinical evaluation means that autism in FXS patients is typically not diagnosed until early school years and that gauging response to therapies for cognitive impairments can suffer from subjective and potentially non-uniform measures of success. The current lack of validated outcome measures hinders the progress of both research and clinical trials, highlighting the value of investing in development of such a measure.

We would like to note here that our research has contributed to the discovery of a novel biomarker based on LIN28-regulated miRNAs that is tested on an animal model. At the current stage of development, our collaborators have identified dysregulation of LIN28 protein and the miRNAs regulated by it. This has been observed in both the brain and peripheral blood of a specific mouse model that mimics the disease state. Extracellular miRNAs have shown great potential as biomarkers in disease as they have been robustly detected in plasma/serum of the blood and are generally highly stable. The very recent

experimental findings indicate statistically significant results with regards to this biomarker in the mouse models that mimicked the disease state as compared to age-matched control mice.

From a therapeutic angle, current ongoing experiments are aimed at determining whether drug treatment in the mouse model with disease state would objectively balance the dysregulated axis and lead to a normalization of the biomarker in the peripheral blood. A subsequent important step would be moving into human FXS patient samples. From a different therapeutic perspective, instead of using a drug to balance the dysregulated axis, current gene-therapy-based experiments are aimed to directly restore the downregulated miRNAs in the mouse model. This would then lead to a future work on developing a potential therapy for human patients with FXS by targeting the dysregulated RNA binding protein-miRNA axis. The significance for this miRNA-based biomarker includes its direct use as a clinical outcome measure in FXS and other ASD clinical trials, and its possible use in diagnosing autism in young children.

Chapter 4

Modeling Synthetic Protein-Protein Interaction Networks in Living Cells

Introduction

In this chapter, our goal is to study biological interaction networks from a different perspective than the one considered in Chapters 2 and 3, by focusing on their dynamic behavior. Our research problem of interest here is to model the dynamics of a protein-protein interaction network in living cells, as an instance of the second broad category of research in systems biology.

Specifically, we investigate a novel strategy for generating intracellular hydrogels, termed iPOLYMER for intracellular Production Of Ligand-Yielded Multivalent EnhanceRs. This particular design, developed by our collaborators (Dr. Takanari Inoue's lab at the School of

Medicine), not only circumvents invasive approaches, such as microinjection, but also enables hydrogel formation inside living cells in a rapidly inducible manner. In the following, we briefly discuss the bioengineering background on hydrogels.

4.1 Bioengineering Background

A hydrogel is a cross-linked polymer network that is capable of absorbing water but does not dissolve in it [56]. An outstanding feature of hydrogels as a material is that their physico-chemical characteristics can be feasibly tuned over a wide range by changing relevant parameters, such as concentrations of polymers and cross-linkers or ambient environments, including temperature and pH. In a biological context, these highly variable properties enable biological hydrogel-like structures to serve versatile roles in living organisms, such as supporting and regulating functional entities [93] as well as lubricating joints [80]. Recent works have found that biological hydrogel-like structures not only exist in extracellular space, but also inside cells [61, 165]. These intracellular hydrogels serve vital functions, such as forming diffusion barriers at the interface of subcellular compartments or nucleating cellular activities [41]. One significant class of intracellular structure that has been related to hydrogels is an RNA granule, which undergoes phase separation-like behavior and dynamic structural rearrangements [12]. Many components of the granule contain low complexity sequences, which actually form hydrogels when purified at higher concentrations. While RNA granules are physiologically important [91], their structural

organization and biological relevance remain uncharacterized. Development of synthetic equivalents of the granules may become an alternative to facilitate our understanding of the relationship between the structure and function of these intracellular hydrogel-like structures.

Synthetic hydrogels have long been of great interest in the field of biomedical engineering, primarily because their physical properties can be designed to achieve a desired objective; e.g., synthetic biomaterials to become a surrogate for damaged tissue [106]. In previous reports, the formation of those gels was successfully controlled in a stimuli-responsive manner [105], providing a wide range of potential applications. Innovations in polymer and protein science have already enabled the development of numerous synthetic hydrogels to be used in clinical practice and bioengineering research [106]. However, the past research has principally focused on extracellular applications, including the design of tissue engineering scaffolds and drug delivery vehicles [73]. Until now, little has been achieved in an effort to generate synthetic hydrogels inside cells, mainly due to the challenging nature of inducing gel formation in intact, living cells. As such, researchers currently resort to either a microinjection of acrylamide gels already formed outside cells [68], or to overexpression of building block molecules whose polymerization cannot be triggered [89]. These reports used the hydrogel as a mechanical probe, described their dynamics in living cells, or evaluated the effects of gel formation on cell survival. However, they have been less successful in directly demonstrating synthetic hydrogels as being biologically functional within living cells, primarily due to the lack of an experimental paradigm that is capable

of forming gels in an inducible manner, with sufficiently fast kinetics to allow monitoring functionality before and after induction.

Another difficulty in studying hydrogel formation inside the cell is the limited toolkits available for gel evaluation. It is not often straightforward to claim the identity of an object observed in the cell as being a gel, owing to limited physical access. Reconstitution of the material in vitro is the most frequently adopted way to address this issue [41], although the conditions adopted in vitro may not necessarily recapitulate the phenomena observed in cells. Comprehensive understanding of the nature of induced hydrogels thus requires a new paradigm, such as a computational model describing the gel formation process.

Given that this research work is a collaborative work, we first aimed to explore the feasibility of iPOLYMER for rapid hydrogel-like network synthesis *in silico*, before our collaborators proceeded with costly and highly time-consuming experimental procedures. Therefore, we developed a physical model for three-component multivalent-multivalent molecular interactions (Fig. 4.1), which led to a rigorous method for computationally implementing iPOLYMER. We then performed kinetic Monte Carlo simulations based on this model.

The proposed model employs a stochastic biochemical reaction system, which contains three types of molecules, FKBP, FRB and rapamycin. These molecules interact according to four reversible reactions (Fig. 4.1) and are subject to random diffusion. The binding sites in FKBP and FRB are labeled as free or bound at each time point, along with the information of binding partners. At a given time, the system may contain a mixture of

FKBP, FRB, and rapamycin molecules, as well as aggregate molecules formed by the mutual binding of these three basic molecules with other larger compound molecules. To model iPOLYMER, we spatially discretize the well-known continuous-space Doi model of stochastic reaction-diffusion [32, 33] and obtain a physically valid approximation based on the reaction-diffusion master equation (RDME) [38, 53, 64, 66, 67]. This leads to a Markov process model that describes the time evolution of the location of each basic or aggregate molecule at a resolution of one voxel in the system. We simulate the resulting process by a stochastic kinetic Monte Carlo algorithm. For our computational analysis, we model the experimental system in a predefined volume of a subcellular size, discretize the system in each spatial direction resulting in a given number of equally sized voxels that satisfy the modeling assumptions and constraints, use experimentally verified kinetic rate values for certain reactions, and plausible values for the kinetic rates of the remaining reactions. In the following, we discuss the details of our methodology and the results and conclusions we obtained from our analyses.

4.2 Model Construction and Simulation

4.2.1 Molecules and Reactions

We consider a biochemical reaction system of volume V that contains three types of molecules: FKBP protein, denoted by an L molecule, FRB protein, denoted by a P

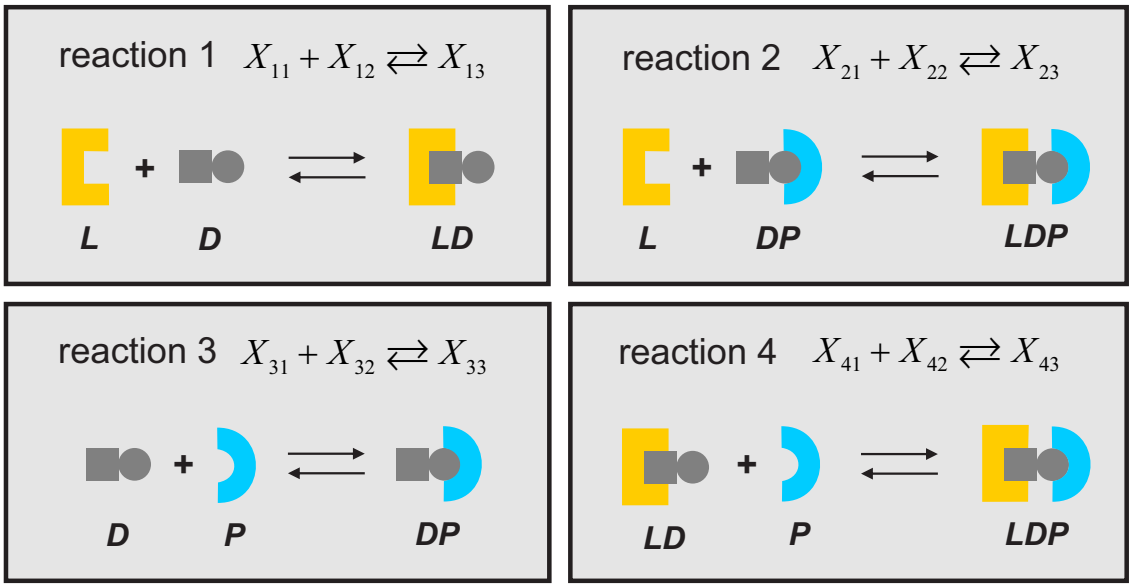

Figure 4.1: The four reactions between molecules L (FKBP), P (FRB) and D (rapamycin), as well as the corresponding reactions among their binding sites.

molecule, and the dimerizing agent rapamycin, denoted by a D molecule. Molecule D comprises two binding sites: an L-binding site that allows D to bind with an L molecule and a P-binding site that allows D to bind with a P molecule. We assume that a molecule L comprises ν_L binding sites for D, whereas a molecule P comprises ν_P binding sites for D. Molecules L and P can bind to each other only when their binding is mediated by molecules D. There are two possibilities: either a molecule L binds on P through a free P-binding site of a D molecule that is bound on L, or a molecule P binds on L through a free L-binding site of a D molecule that is bound on P.

The previous molecules interact according to the four reactions depicted in Fig. 4.1, which correspond to the following reactions among binding sites:[1]

[1]Note that Fig. 4.1 does not show the fact that molecules L and P comprise multiple binding sites for molecules D. Moreover, it does not show the fact that each reactant molecule L, D, and P can be part of a larger molecule (i.e., an aggregate).

$$X_{11} + X_{12} \rightleftarrows X_{13} \text{ (binding/unbinding of a free site on } L \text{ with a free site on an unbound } D) \tag{4.1}$$

$$X_{21} + X_{22} \rightleftarrows X_{23} \text{ (binding/unbinding of a free site on } L \text{ with a free site on a bound } D) \tag{4.2}$$

$$X_{31} + X_{32} \rightleftarrows X_{33} \text{ (binding/unbinding of a free site on } P \text{ with a free site on an unbound } D) \tag{4.3}$$

$$X_{41} + X_{42} \rightleftarrows X_{43} \text{ (binding/unbinding of a free site on } P \text{ with a free site on a bound } D) \tag{4.4}$$

In these reactions, the X_{mn}'s denote sets of single and paired binding sites, where:

- X_{11} is the set of free D-binding sites on molecules L
- X_{12} is the set of free L-binding sites on unbound molecules D
- X_{13} is the set of bound pairs of binding sites between molecules L and D, in which the P-binding site on the D molecule is free
- $X_{21} = X_{11}$
- X_{22} is the set of free L-binding sites on molecules D bound on P
- X_{23} is the set of bound pairs of binding sites between molecules L and D, in which

the P-binding site on the D molecule is bound

- X_{31} is the set of free P-binding sites on unbound molecules D
- X_{32} is the set of free D-binding sites on molecules P
- X_{33} is the set of bound pairs of binding sites between molecules P and D, in which the L-binding site on the D molecule is free
- X_{41} is the set of free P-binding sites on molecules D bound on L
- $X_{42} = X_{32}$
- X_{43} is the set of bound pairs of binding sites between molecules P and D, in which the L-binding site on the D molecule is also bound

Note that the first subscript m in X_{mn} is used to denote the reaction, whereas the second subscript n is used to denote the order of the "species" associated with each reaction (e.g., X_{11}, X_{12} and X_{13} are the first, second, and third "species" associated with reaction 1).

The previous forward reactions are second-order, whereas the reverse reactions are first-order. When a forward reaction is about to occur, the reactive components (i.e., the individual binding sites) are first identified. Upon occurrence of the reaction, these binding sites are removed from the corresponding reactive components and the resulting bound pair of binding sites is added to the product. On the other hand, when a reverse reaction is about to occur, the reactive component (i.e., the individual bound pair of binding sites) is first identified. Upon occurrence of the reaction, the pair of binding sites is removed from the

reactive component and the resulting individual binding sites are added to the corresponding products.

As a result of reactions (4.1)-(4.4), the system may contain a mixture of the three basic molecules L, P, and D, as well as aggregate molecules resulting from their mutual binding. In addition, reactions (4.1)-(4.4) occur randomly, whereas molecules present in the reaction system will randomly diffuse within the volume V. In the following, we develop a stochastic reaction-diffusion computational model for these processes.

4.2.2 Available Models

Three reaction-diffusion physical models have been proposed in the literature to study stochastic biochemical reactions in cellular systems. These are the Doi model, the Smoluchowski diffusion limited reaction model, and a model based on the *reaction-diffusion master equation* (RDME) [66].

In the Doi model [32, 33], molecules are represented as points inside a three-dimensional volume. First- and second-order reactions are assumed to occur with fixed probability rates (probabilities per unit time). However, a second-order reaction occurs only when its two reactants are separated by *less* than a specified "reaction radius" r. The product of a second-order reaction of the form $A+B \rightarrow C$ is usually placed midway between the two reactants, whereas the products of a reverse first-order reaction $C \rightarrow A + B$ are placed at a distance that is not appreciably larger than r. Finally, diffusion of molecules is modeled by independent Brownian motions.

The Smoluchowski model [72] is similar to the Doi model. However, this model assumes that second-order reactions may occur either instantaneously or when its two reactants are separated *exactly* by the reaction radius r.

The RDME model can be interpreted as an extension of the non-spatial chemical master equation model for stochastic chemical kinetics [42, 67, 158]. In this case, molecules are represented as points inside a three-dimensional volume. The volume is partitioned by a rectangular mesh into voxels and diffusion is modeled as a jump of a molecule from its current voxel into one of its neighboring voxels. First- and second-order reactions occur within a voxel with fixed probability rates and independently from the reactions in any other voxel. Moreover, second-order reactions occur in a voxel under the assumption of well-mixed reactants within the voxel. This is justified by assuming that the mesh spacing is appreciably larger than the reaction radius and that the reactants become well-mixed at an appreciably faster timescale than that of the occurrence of second-order reactions.

It turns out that the Doi model offers comparable accuracy to the Smoluchowski model [66, 94]. For this reason, we assume that the physical properties of the biological system at hand can be modeled sufficiently well by the Doi model. Note that the Doi model is continuous in space and cannot be used for computational analysis. To develop a computational approach to our problem, we need to discretize this model. It turns out that the RDME model can be interpreted as a formal discrete-space approximation of the Doi (or the Smoluchowski) model [38, 53, 64, 66].

We should mention here that when the reaction system contains second-order reactions

(which is true in our case), using the RDME model as an approximation to the Doi model can be problematic. This is due to the fact that, in the limit, as the spacing of the three-dimensional mesh approaches zero, second-order reactions are lost in the RDME model (i.e., these reactions never occur) [53, 65]. The main reason for this problem is that, for very small mesh spacings, the reactants within a voxel may not become well-mixed before a reaction occurs. As a consequence, the error in approximating the Doi model by the RDME model cannot be made arbitrarily small.

A variant of the RDME model, called *convergent* RDME (CRDME), has been proposed in [66] to address the previous issue. This model promptly converges to the Doi model in the limit as the mesh spacing tends to zero. Note however that the introduction of the CRDME model in no way invalidates the use of the standard RDME model as an approximation to the Doi model. The work presented in [66] provides choices for the appropriate mesh spacing and the underlying parameters for which the RDME is considered a physically valid approximation to the CRDME and Doi models.

4.2.3 RDME-based Approach

By following standard RDME modeling steps, we first employ a three-dimensional rectangular mesh to partition the volume V of the biochemical reaction system at hand into N^3 equally-sized voxels. We consider a cubic volume $V = [0, S] \times [0, S] \times [0, S]$, which implies a uniform mesh spacing of $s = S/N$. In this case, the volume of each voxel is given by $V_0 = s^3$. We assume that, within each voxel, molecules are well-mixed (i.e., uniformly

distributed). As a consequence, the position of each molecule is known only to the scale of one voxel. Note that molecules are considered to be points in the three-dimensional space and are, therefore, dimensionless. This is not only true for the three basic molecules L, P and D, but also for their forming aggregates.

We model diffusion as a first-order reaction that results in a jump of a molecule from its current voxel to a non-diagonal nearest-neighbor voxel. In the following, we denote by d_{ij} the probability rate of the reaction that models diffusion of a specific molecule within a voxel i to a voxel j. We assume that every molecular species in voxel i is characterized by the same probability rate of diffusion. We also assume that diffusion can be characterized by the same probability rate irrespective of the particular voxel i and its nearest-neighbor voxel j. To ensure that, when none of the reactions (4.1)-(4.4) occur, we can correctly recover the diffusion of individual molecules in the limit as $s \to 0$, we must set [64, 67]

$$d_{ij} = \begin{cases} d/s^2, & \text{if } j \text{ is a non-diagonal nearest-neighbor of } i \\ 0, & \text{otherwise,} \end{cases}$$

where d is the diffusion coefficient.

Note that, as aggregates start to form, the reaction system will contain a large number of different types of molecules, which are expected to be characterized by different diffusion coefficients. Using the Stokes-Einstein formula to calculate the values of these coefficients is not possible, since this equation is valid for spherical particles [74], whereas the structure of relatively small compound molecules are mainly tree-like. Therefore, obtaining reliable

values for the diffusion coefficients of each individual compound molecule is an extremely difficult problem. For this reason, it is not practical to assign different probability rates of diffusion to each type of molecule in the system. However, the proposed procedure can be modified to accommodate different rates if such rates become available.

Our previous choice for the probability rate of diffusion implies that our computational model will take less time to reach steady state since, in reality, larger molecules would diffuse more slowly than smaller molecules. Moreover, there are certain other factors that contribute to the difference in time scale between our computational and experimental findings, which we review in the Discussion section.

In the Doi model, the forward reactions (4.1)-(4.4) occur only if the associated reactants are within the reaction radius r. Let λ_m^+, $m = 1, 2, 3, 4$, be the probability rates of these reactions. Moreover, let λ_m^-, $m = 1, 2, 3, 4$, be the probability rates of the corresponding reverse reactions. In the following, we assume that the mesh spacing s is appreciably larger than the reaction radius r (i.e., we assume that $s \gg r$). Intuitively speaking, this condition guarantees the two main premises underlying the RDME model: most forward reactions within a voxel will be between reactants within the same voxel whereas the products of most reverse reactions will be placed within the same voxel.

If, in the RDME model, we choose the probability rate κ_d of a diffusion reaction to be given by

$$\kappa_d = \frac{d}{s^2},$$

and the probability rates κ_m^+ and κ_m^- of the binding and unbinding reactions (4.1)-(4.4) to

be given by

$$\kappa_m^+ = \frac{4\pi r^3}{3s^3}\lambda_m^+ \quad \text{and} \quad \kappa_m^- = \lambda_m^-, \quad \text{for } m = 1, 2, 3, 4, \tag{4.5}$$

then the RDME model may be interpreted as an asymptotic (as $r/s \to 0$) approximation of the CRDME model (and thus of the Doi model). The error in approximation depends on the values of r, λ and d. It turns out that, for a fixed ratio r/s, smaller values of $r\sqrt{\lambda/d}$ result in a better approximation of the CRDME and Doi models by the RDME model [66].

An important practical issue to consider here is choosing an appropriate value for the mesh spacing s. On one hand, the previous results force us to take $s \gg r$ in order to use the RDME model as a reasonable approximation to the CRDME model. On the other hand, we must take s small enough so that $s \ll S$. This will ensure that discretization of the continuous problem will be fine enough to guarantee sufficiently accurate approximation of system behavior, both in terms of the binding/unbinding reactions within a voxel, as well as in terms of molecular diffusion [65].

Recall now that a fundamental assumption associated with the RDME model is that molecules within a voxel are well-mixed before one of the binding reactions (4.1)-(4.4) occurs. To make sure that this is the case, we must assume that the timescale τ_d for the reactants to become well-mixed due to diffusion is appreciably smaller than the timescale τ_r of each of the binding reactions (4.1)-(4.4) to occur; i.e., we must have that $\tau_d \ll \tau_r$. As we discussed previously, the probability rate of diffusion of a selected molecule is given by d/s^2. This implies that the timescale of diffusion will be approximately equal to s^2/d; i.e., $\tau_d \simeq s^2/d$ [65, 67]. On the other hand, the probability rates of the binding reactions

in a well-mixed voxel are given by Eq. (4.5). In this case, the timescale τ_r of the binding reactions will approximately be equal to $1/\max_m\{\kappa_m^+\}$, in which case

$$\tau_r \simeq \frac{3s^3}{4\pi r^3 \max_m\{\lambda_m^+\}}.$$

As a consequence, and in order for $\tau_d \ll \tau_r$, we must have that

$$s \gg \frac{4\pi}{3}\frac{r^3}{d}\max_m\{\lambda_m^+\}.$$

Therefore, proper use of the RDME model requires that we employ a mesh spacing that satisfies the following inequality

$$s' \ll s \ll s'', \tag{4.6}$$

where

$$s' = \max\left\{r, \frac{4\pi}{3}\frac{r^3}{d}\max_m\{\lambda_m^+\}\right\} \qquad \text{and} \qquad s'' = \sqrt[3]{V}.$$

Note that the first-order unbinding reactions (4.1)-(4.4) do not restrict the mesh spacing since they represent internal molecular events which do not require the molecules within a voxel to be well-mixed.

Condition (4.6) underlies the required assumption that the molecules within a voxel are well-mixed before any binding reaction occurs. It moreover leads to the assumption that the binding/unbinding reactions within a voxel will occur independently from the reactions within any other voxel. This is due to the fact that condition (4.6) implies that the vast

majority of the reactants and products of the binding/unbinding reactions in a voxel will be inside that voxel. Note also that the diffusion reactions within a given voxel are mutually independent and independent from the binding/unbinding reactions within the voxel. Moreover, these reactions are independent from the diffusion and binding/unbinding reactions within any other voxel. As a consequence, we can approximately partition the biochemical reaction system under consideration with volume V into V/s^3 statistically independent biochemical reaction subsystems of equal volume s^3. Each subsystem comprises a number of binding/unbinding reactions and a set of mutually independent diffusion reactions that are also independent of the binding/unbinding reactions.

Let us now denote by $n_i(t)$ the total number of molecules within voxel i at time t. By following the exact algorithm of Gillespie [45], the probability that the next diffusion reaction will occur at time $t+\tau+dt$ within voxel i is governed by an exponential distribution with rate parameter $\kappa_d n_i(t)$, whereas the probability of a specific molecule to be diffused in any one of the six possible directions is given by

$$\frac{\kappa_d}{6\kappa_d n_i(t)} = \frac{1}{6n_i(t)}.$$

Let us also denote by $n^+_{i,m}(t)$ the total number of pairs of binding sites within voxel i at time t, with each pair consisting of sites located on distinct molecules, which can potentially react through the m-th binding reaction (4.1)-(4.4). Moreover, let $n^-_{i,m}(t)$ be the total number of pairs of bound sites within voxel i at time t that can potentially unbound

through the m-th unbinding reaction (4.1)-(4.4). In this case, the probability that the next binding/unbinding reaction will occur at time $t + \tau + dt$ within voxel i is governed by an exponential distribution with rate parameter

$$\sum_{m'=1}^{4} \Big[\kappa^{+}_{m'} n^{+}_{i,m'}(t) + \kappa^{-}_{m'} n^{-}_{i,m'}(t) \Big],$$

whereas the probabilities of a specific binding/unbinding reaction to occur is given by

$$\kappa^{+}_{m} n^{+}_{i,m}(t) \Big[\sum_{m'=1}^{4} \kappa^{+}_{m'} n^{+}_{i,m'}(t) + \kappa^{-}_{m'} n^{-}_{i,m'}(t) \Big]^{-1}, \quad \text{for binding,}$$

and

$$\kappa^{-}_{m} n^{-}_{i,m}(t) \Big[\sum_{m'=1}^{4} \kappa^{+}_{m'} n^{+}_{i,m'}(t) + \kappa^{-}_{m'} n^{-}_{i,m'}(t) \Big]^{-1}, \quad \text{for unbinding.}$$

When a diffusion reaction is about to occur at time $t + \tau + dt$, the individual molecule to be diffused is first identified by choosing it uniformly, with probability $1/n_i(t)$, among all possible molecules within voxel i. Subsequently, the direction of diffusion is identified by choosing it uniformly among all possible directions with probability $1/6$. Upon occurrence of the reaction, the binding sites of the diffused molecule are relabeled to indicate their jump from voxel i to the new voxel j.

On the other hand, when a binding reaction m is about to occur at time $t + \tau + dt$, the individual pair of binding sites are first identified by choosing them uniformly, with probability $1/n^{+}_{i,m}(t)$, among all possible pairs of binding sites. Upon occurrence of the

reaction, these binding sites are removed from the corresponding reactive components and the resulting bound pair of binding sites is added to the corresponding product. On the other hand, when an unbinding reaction m is about to occur at time $t + \tau + dt$, the individual bound pair of binding sites is first chosen uniformly, with probability $1/n^-_{i,m}(t)$, among all possible bound pairs. Upon occurrence of the reaction, the pair of binding sites is removed from the corresponding reactive component and the resulting individual binding sites are added to the corresponding products for the given reaction.

4.2.4 Choosing Kinetic Values

The following kinetic rate values for reactions (4.1) & (4.4) have been experimentally specified in [7]:

$$k_1^+ = 5.8 \times 10^6 \mathrm{M}^{-1}\mathrm{sec}^{-1}$$

$$k_1^- = 1.6 \times 10^{-3}\mathrm{sec}^{-1}$$

$$k_4^+ = 1.7 \times 10^6 \mathrm{M}^{-1}\mathrm{sec}^{-1}$$

$$k_4^- = 1.9 \times 10^{-2}\mathrm{sec}^{-1}.$$

Due to experimental difficulties however only values for the dissociation constants of reactions (4.2) & (4.3) have been provided in [7], given by

$$K_2 = k_2^-/k_2^+ \simeq 1 \times 10^{-13}\mathrm{M}$$

$$K_3 = k_3^-/k_3^+ \simeq 23 \times 10^{-6}\mathrm{M}\,.$$

Since our model requires kinetic rates for all reactions, we can obtain plausible values for the kinetic rates of reaction 2 by assuming that the rate of FKBP binding to an FRB-rapamycin complex is α times faster than the rate of FKBP binding to a free rapamycin molecule, whereas the rate of FKBP unbinding from an FRB-rapamycin complex is α times slower than the rate of FKBP unbinding from rapamycin alone. In this case,

$$K_2 = \frac{k_2^-}{k_2^+} = \frac{k_1^-/\alpha}{\alpha k_1^+} = \frac{1}{\alpha^2}\frac{k_1^-}{k_1^+},$$

which implies that

$$\alpha = \sqrt{\frac{1}{K_2}\frac{k_1^-}{k_1^+}} = \sqrt{\frac{1}{100\text{fM}}\frac{1.6 \times 10^{-3}\text{sec}^{-1}}{5.8 \times 10^{6}\text{M}^{-1}\text{sec}^{-1}}} = 52.52.$$

As a consequence, we obtain

$$k_2^+ = 3.1 \times 10^{8}\text{M}^{-1}\text{sec}^{-1}$$

$$k_2^- = 3.1 \times 10^{-5}\text{sec}^{-1}\ .$$

Similarly, we can obtain plausible values for the kinetic rates of reaction 3 by assuming that the rate of FRB binding to a free rapamycin molecule is α times slower than the rate of FRB binding on an FKBP-rapamycin complex, whereas the rate of FRB unbinding from a free rapamycin molecule is α times faster than the rate of FRB unbinding from an FKBP-

rapamycin-FRB complex. In this case,

$$K_3 = \frac{k_3^-}{k_3^+} = \frac{\alpha k_4^-}{k_4^+/\alpha} = \alpha^2 \frac{k_4^-}{k_4^+},$$

which implies that

$$\alpha = \sqrt{K_3 \frac{k_4^+}{k_4^-}} = \sqrt{23 \times 10^{-6}\mathbf{M} \frac{1.7 \times 10^6 \mathbf{M}^{-1}\text{sec}^{-1}}{1.9 \times 10^{-2}\text{sec}^{-1}}} = 45.36.$$

As a consequence, we obtain

$$k_3^+ = 3.8 \times 10^4 \mathbf{M}^{-1}\text{sec}^{-1}$$

$$k_3^- = 8.6 \times 10^{-1}\text{sec}^{-1} \ .$$

Since our model utilizes probability rates, the previous values of the kinetic rates must be translated to probability rates. This can be done by noting the following relationship between the probability rate κ of a reaction and its kinetic rate k:

$$\kappa^+ = \frac{k^+}{AV_0} \qquad \text{and} \qquad \kappa^- = k^-, \tag{4.7}$$

where $A = 6.022 \times 10^{23}$ mol^{-1} is Avogadro's number and V_0 is the voxel volume measured in litres. In this case, κ is measured in sec^{-1}. As a consequence of Eqs. (4.5) & (4.7), we

have for the forward reactions that

$$\lambda^{+} = \frac{3s^3 k^{+}}{4\pi r^3 A V_0} \text{ sec}^{-1},$$

whereas, for the reverse reactions, we have that

$$\lambda^{-} = k^{-} \text{ sec}^{-1}.$$

By employing these formulas, we can use the previous values for the kinetic rates as inputs to our model.

4.2.5 Simulation via Kinetic Monte Carlo

The previously discussed RDME-based approach leads to a stochastic Markovian process that describes the time evolution of the location of each basic or aggregate molecule within volume V at a resolution of one voxel. To simulate this process, we could use standard kinetic Monte Carlo which leads to the following simulation algorithm:

Exact Simulation Algorithm

1. Specify values for the following parameters:
   - V (system volume)
   - s (mesh spacing)
   - ν_L (valence number of L; i.e., number of binding sites on molecule L for D)

- ν_P (valence number of P; i.e., number of binding sites on molecule P for D)
- N_L (initial number of L molecules)
- N_D (initial number of D molecules)
- N_P (initial number of P molecules)
- r (reaction radius)
- d (diffusion coefficient)
- λ_m^+, $m = 1, 2, 3, 4$ (physical probability rates of binding reactions)
- λ_m^-, $m = 1, 2, 3, 4$ (physical probability rates of unbinding reactions)
- t_{max} (simulation time).

2. Compute the probability rates $\kappa_d = d/s^2$, $\kappa_m^+ = (4\pi r^3 \lambda_m^+)/(3s^3)$, and $\kappa_m^- = \lambda_m^-$ of the RDME model, for $m = 1, 2, 3, 4$.

3. Initialize the molecular population by independently placing each molecule L in a voxel, uniformly chosen from all possible voxels. Repeat this process for molecules D and P. Set $t = 0$.

4. Choose a voxel i with probability s^3/V uniformly among all possible voxels and compute the total number $n_i(t)$ of molecules present in that voxel. Moreover, for each $m = 1, 2, 3, 4$, compute the total number $n_{i,m}^+(t)$ of pairs of binding sites, with each pair consisting of sites located on distinct molecules, which can potentially react through the m-th binding reaction (4.1)-(4.4), and compute the total number $n_{i,m}^-(t)$

of pairs of bound sites that can potentially unbound through the m-th unbinding reaction (4.1)-(4.4).

5. Determine the time of the next diffusion reaction within voxel i by drawing a sample t_d from an exponential distribution with rate parameter $\kappa_d n_i(t)$. In addition, determine the time of the next binding/unbinding reaction within voxel i by drawing a sample t_r from an exponential distribution with rate parameter $\sum_{m'=1}^{4}[\kappa_{m'}^{+} n_{i,m'}^{+}(t) + \kappa_{m'}^{-} n_{i,m'}^{-}(t)]$.

6. If $t_d \leq t_r$, determine the direction of diffusion with uniform probability $1/6$ over all six possible directions as well as the particular molecule to be diffused with uniform probability $1/n_i(t)$ among all possible molecules in voxel i. Move the selected molecule from voxel i to the new voxel j and relabel the binding sites of the diffused molecule to indicate their jump from voxel i to voxel j. If $t_d < t_r$, set $t = t + t_d$ and go to step 4.

7. Determine the binding/unbinding reaction to occur at time t_r by drawing a sample from the probability mass

$$\left\{ \kappa_m^{+} n_{i,m}^{+}(t) \Big[\sum_{m'=1}^{4} \kappa_{m'}^{+} n_{i,m'}^{+}(t) + \kappa_{m'}^{-} n_{i,m'}^{-}(t) \Big]^{-1}, \right.$$
$$\left. \kappa_m^{-} n_{i,m}^{-}(t) \Big[\sum_{m'=1}^{4} \kappa_{m'}^{+} n_{i,m'}^{+}(t) + \kappa_{m'}^{-} n_{i,m'}^{-}(t) \Big]^{-1}, m = 1, 2, 3, 4 \right\}.$$

If the m-th binding reaction is drawn, identify the individual pair of binding sites

to react by choosing them uniformly, with probability $1/n^{+}_{i,m}(t)$, among all possible pairs of binding sites. If the m-th unbinding reaction is drawn, uniformly choose the individual bound pair of sites to react, with probability $1/n^{-}_{i,m}(t)$, among all possible bound pairs. Appropriately adjust the corresponding reactive and product components to reflect the occurrence of the reaction. Set $t = t + t_r$ and go to step 4.

8. Terminate the algorithm if $t > t_{\text{max}}$.

It turns out that exact spatial stochastic simulations can be costly. Essentially, this is caused by the fact that refinement of the discretized spatial domain, together with the relatively large number of molecules in the reaction system, give rise to a large number of diffusive events between voxels. As a consequence, the stochastic simulation of the reaction-diffusion system eventually becomes dominated by diffusive transfer events that occur much more frequently than chemical reactions.

To address the previous issue, we utilize an approximation technique proposed in [121], which is built on the multi-particle lattice gas automaton model presented in [22]. The performance of this technique has been validated in [31] and leads to the following simulation algorithm:

Approximate Simulation Algorithm

1. Specify values for the following parameters:
   - V (system volume)
   - s (mesh spacing)

- ν_L (valence number of L; i.e., number of binding sites on molecule L for D)
- ν_P (valence number of P; i.e., number of binding sites on molecule P for D)
- N_L (initial number of L molecules)
- N_D (initial number of D molecules)
- N_P (initial number of P molecules)
- r (reaction radius)
- d (diffusion coefficient)
- λ_m^+, $m = 1, 2, 3, 4$ (physical probability rates of binding reactions)
- λ_m^-, $m = 1, 2, 3, 4$ (physical probability rates of unbinding reactions)
- t_{max} (simulation time).

2. Compute the probability rates $\kappa_m^+ = (4\pi r^3 \lambda_m^+)/(3s^3)$, and $\kappa_m^- = \lambda_m^-$ of the RDME model, for $m = 1, 2, 3, 4$.

3. Initialize the molecular population by independently placing each molecule L in a voxel, uniformly chosen from all possible voxels. Repeat this process for molecules D and P.

4. Set $t = 0$, $\tau_d = s^2/6d$, $n_d = 1$, and $N = \sqrt[3]{V}/s$.

5. For each voxel $i = 1, 2, \ldots, N^3$ and for each $m = 1, 2, 3, 4$, compute the total number $n_{i,m}^+(t)$ of pairs of binding sites that can potentially react through the m-th binding reaction (4.1)-(4.4), with each pair consisting of sites located on distinct molecules.

Moreover, compute the total number $n^-_{i,m}(t)$ of pairs of bound sites that can potentially unbound through the m-th unbinding reaction (4.1)-(4.4).

6. Determine the time $t + \tau_r$ of the next binding/unbinding reaction among all voxels by drawing a sample τ_r from the exponential distribution with rate parameter $\sum_{i=1}^{N^3} \sum_{m'=1}^{4} [\kappa^+_{m'} n^+_{i,m'}(t) + \kappa^-_{m'} n^-_{i,m'}(t)]$.

7. If $t + \tau_r > t_{\text{max}}$ terminate the simulation.

8. Determine which binding/unbinding reaction will occur at time $t + \tau_r$ by drawing a sample from the probability mass

$$\Bigg\{ \kappa^+_m n^+_{i,m}(t) \Big[\sum_{i=1}^{N^3} \sum_{m'=1}^{4} \kappa^+_{m'} n^+_{i,m'}(t) + \kappa^-_{m'} n^-_{i,m'}(t) \Big], \; \kappa^-_m n^-_{i,m}(t) \Big[\sum_{i=1}^{N^3} \sum_{m'=1}^{4} \kappa^+_{m'} n^+_{i,m'}(t) + \kappa^-_{m'} n^-_{i,m'}(t) \Big], \; m = 1,2,3,4, i = 1,2,\ldots,N^3 \Bigg\}.$$

If the (m,i)-th binding reaction (the m-th binding reaction in voxel i) occurs, identify the individual pair of binding sites to react, by choosing them uniformly with probability $1/n^+_{i,m}(t)$ among all possible pairs of binding sites. If the (m,i)-th unbinding reaction (the m-th unbinding reaction in voxel i) occurs, identify the individual pair of bound sites to react, by choosing them uniformly with probability $1/n^-_{i,m}(t)$ among all possible bound pairs. Appropriately adjust the corresponding reactive and product components to reflect the occurrence of the particular reaction.

9. Increment t by τ_r. If $t < n_d \tau_d$, go to Step 5.

10. For voxels $i = 1, 2, \ldots, N^3$, diffuse each species in each voxel by determining the direction of diffusion with uniform probability $1/6$ over all six possible directions. Move a selected molecule from voxel i to the new voxel j and relabel the binding sites of the diffused molecule to indicate its jump from voxel i to voxel j.

11. Increment n_d by 1 and set $t = n_d \tau_d$.

12. If $t > t_{\text{max}}$ terminate the simulation. Otherwise, go to Step 5.

It is noteworthy to mention here that the reason why several distinct models have been proposed in the literature to study stochastic and spatial effects in biochemical reaction systems (e.g., see [31, 46, 148]) is because no single model is currently capable of efficiently coping with the broad range of spatial, temporal and concentration scales commonly found in biochemical reaction networks. For this reason, models such as the one discussed in this research, may represent a plausible approach that yields a compromise between computational efficiency as well as spatial and stochastic accuracy.

4.2.6 Size Distribution of Molecular Aggregates

To study the dynamic formation of molecular aggregates, we need to define the size of a given molecule in the system at a given time. This will enable us to calculate and track the size distribution of molecules in the system as a function of time. Moreover, this information will allow us to study properties of molecular aggregation in terms of the concentration and valence numbers of individual molecules initially present in the system.

Understanding the evolutionary behavior of size distribution of molecules in the system was important to our collaborators in their effort to design appropriate experiments for hydrogel-like network synthesis.

Here, we define the size of a particular molecular aggregate to be the *net* number of L and P molecules contained in the aggregate (as a consequence, individual L and P molecules are characterized by unit size). We do not include D molecules in the definition of an aggregate's size since this can be misleading. To see why this is true, let us consider two aggregates, one comprising 50 L and P molecules as well as 200 D molecules, and the other 70 L and P molecules as well as 50 D molecules. If we include molecules D in the definition of size, then the size of the first aggregate will be 250 whereas the size of the second aggregate will be 120. However, the physical size of a D molecule is negligible when compared to the physical size of an L or a P molecule (Fig. 4.2). As a consequence, we expect that the size of the first aggregate will be smaller than that of the second aggregate, since the former contains a smaller number of L and P molecules, which are the dominant determinants of an aggregate's size.

We can summarize the evolutionary nature of aggregate formation as a function of time by employing the *size distribution* of molecular aggregates at different time points. We define this distribution by

$$\mathcal{S}_t(\sigma) = \frac{\mathrm{E}[N_t(\sigma)]}{\mathrm{E}[N_t]}, \quad \text{for } \sigma = 1, 2, \ldots,$$

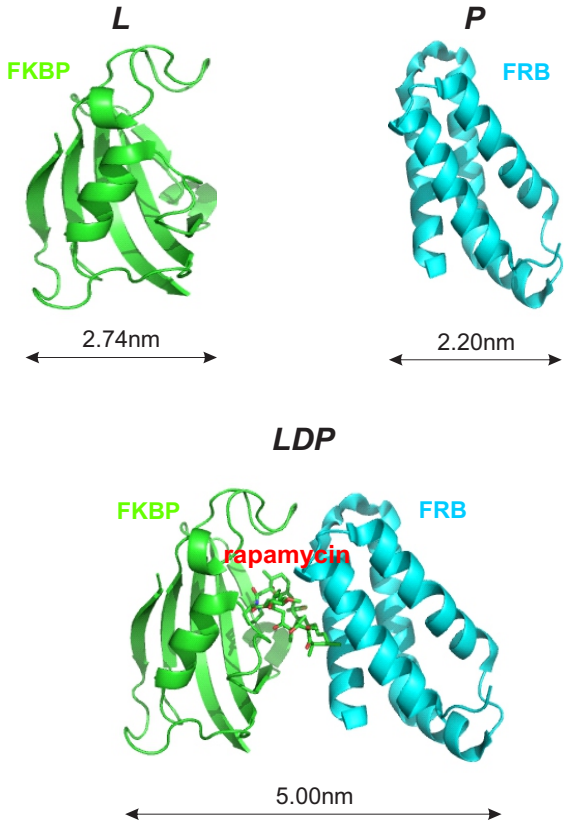

Figure 4.2: Physical sizes of L (FKBP) and P (FRB) molecules as well as of the LDP (FKBP-rapamycin-FRB) complex.

where $N_t(\sigma)$ is the number of aggregates with size σ at time t, N_t is the *net* number of aggregates at time t, and $\mathrm{E}[\cdot]$ denotes expectation.[2] Unfortunately, we cannot calculate the size distribution analytically. We can however estimate it computationally via Monte Carlo. We can do this by employing K kinetic Monte Carlo simulation runs of the RDME-based model and by computing

$$\widehat{\mathcal{S}}_t(\sigma) = \frac{\sum_{k=1}^{K} N_t^{(k)}(\sigma)}{\sum_{k=1}^{K} \sum_{\sigma' \geq 1} N_t^{(k)}(\sigma')}, \tag{4.8}$$

[2]Recall that aggregate formation is a random process.

where $N_t^{(k)}(\sigma)$ is the number of aggregates with size σ at time t, obtained from the k-th simulation run.

4.3 RDME-based Simulation Results

To illustrate the potential of the previous RDME-based computational model, we now provide a number of simulation results we obtained with our model. We use the following choices for the parameter values:

$S = 1\mu$m, which corresponds to $V = 10^{-15}$ l

$s = 0.25\mu$m, which corresponds to $V_0 = 1.5625 \times 10^{-17}$ l

$\nu_{_L} = 1, 2, 3, 4, 5$

$\nu_{_P} = 1, 2, 3, 4, 5$

$N_{_L} = 200$

$N_{_P} = 200$

$N_{_D} = \max\{\nu_{_L}, \nu_{_P}\} \times \max\{N_{_L}, N_{_P}\}$

$r = 1$nm

$d = 10\mu\text{m}^2\text{sec}^{-1}$

$k_1^+ = 5.8 \times 10^6\text{M}^{-1}\text{sec}^{-1}$

$$k_1^- = 1.6 \times 10^{-3}\text{sec}^{-1}$$

$$k_2^+ = 3.1 \times 10^{8}\text{M}^{-1}\text{sec}^{-1}$$

$$k_2^- = 3.1 \times 10^{-5}\text{sec}^{-1}$$

$$k_3^+ = 3.8 \times 10^{4}\text{M}^{-1}\text{sec}^{-1}$$

$$k_3^- = 8.6 \times 10^{-1}\text{sec}^{-1}$$

$$k_4^+ = 1.7 \times 10^{6}\text{M}^{-1}\text{sec}^{-1}$$

$$k_4^- = 1.9 \times 10^{-2}\text{sec}^{-1}$$

$$t_{\text{max}} = 3\text{sec}$$

In order to guarantee that there are enough D molecules to bind all available L- and P-binding sites, we set the initial number of D molecules to be $\max\{\nu_L, \nu_P\} \times \max\{N_L, N_P\}$. Note that $s' = 0.05\mu$m and $s'' = 1\mu$m, in which case, s satisfies the inequalities in Eq. (4.6).

Kinetic Monte Carlo simulation of the previous RDME-based model using the Approximate Simulation Algorithm resulted in the accompanied Videos 1-5, corresponding to $\nu_L = \nu_P = 1, 2, 3, 4, 5$. These videos depict samples of the dynamic behavior of aggregate formation as a function of time over a 2-D planar projection of the 3-D volume V.[3] Due to lack of detailed information about the 3-D geometry of molecular aggregates, we represent an aggregate as a sphere in 3-D, and thus as a disk in 2-D, whose diameter is proportional to the aggregate's size with proportionality constant 1.25×10^{-3}.

[3]With our values for V and s, the volume is partitioned into $V/s^3 = 64$ voxels. As a consequence, its 2-D planar projection is partitioned into 16 squares.

Recall that the RDME-based model specifies the position of each molecule at a resolution of one voxel. As a consequence, it is not possible to visually distinguish molecules within a particular voxel and, therefore, within the square defined by the voxel's 2-D planar projection. To address this issue, and for the purpose of visualization, we arbitrarily shift each molecule within a square uniformly over that square.

It is clear from the results visualized in Videos 1-5 that, at the beginning of each simulation, individual unbound molecules are scattered uniformly within the volume and thus within the 2-D planar projection. As time proceeds however larger molecules (i.e., aggregates comprising many L and P molecules) quickly emerge in the videos that correspond to higher valence numbers, whereas a single large aggregate dominates the population at the end of the simulation.

In Figs. 4.3-4.7, we depict log plots of estimated size distributions of aggregates, which are formed at six equally-spaced time points between 0 and 3 sec, corresponding to valence numbers $\nu_L = \nu_P = 1, 2, 3, 4, 5$. We obtained these distributions from Eq. (4.8) by employing $K = 100$ kinetic Monte Carlo simulation runs of the RDME-based model. These distributions are binned into groups of 10 consecutive sizes. The results summarize the dynamic formation of aggregates from smaller molecules and delineate the fact that, when the valence numbers of the individual molecules are sufficiently high, complex aggregate molecules appear in the course of time. Note that, at $t = 0$, the size distribution takes value 1 at size 1. This is a consequence of the fact that the reaction system contains only individual molecules. At later times, the size distribution spreads over larger sizes, reflecting the

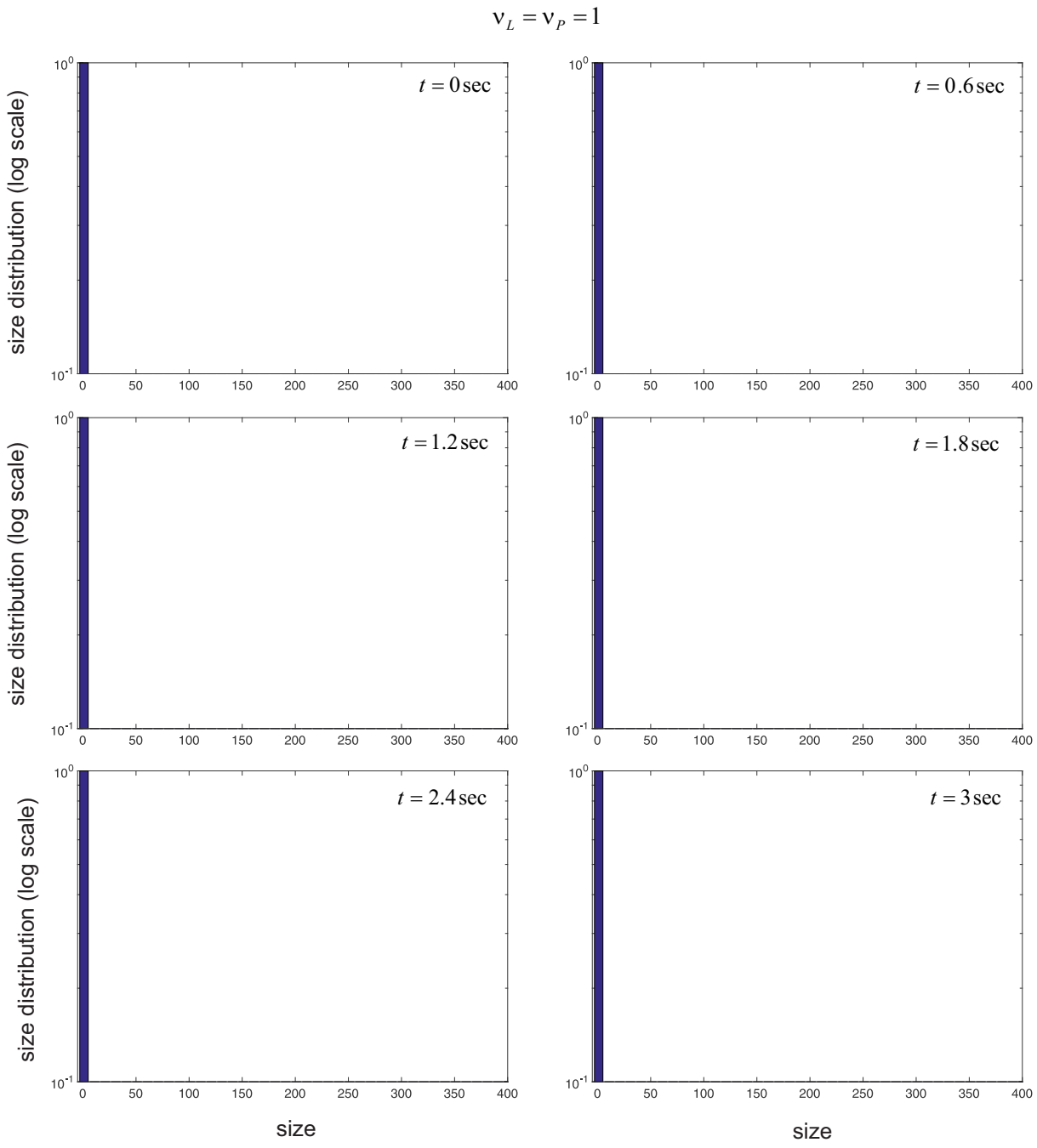

Figure 4.3: Estimated size distributions of molecular aggregates at different time points when $\nu_L = \nu_P = 1$.

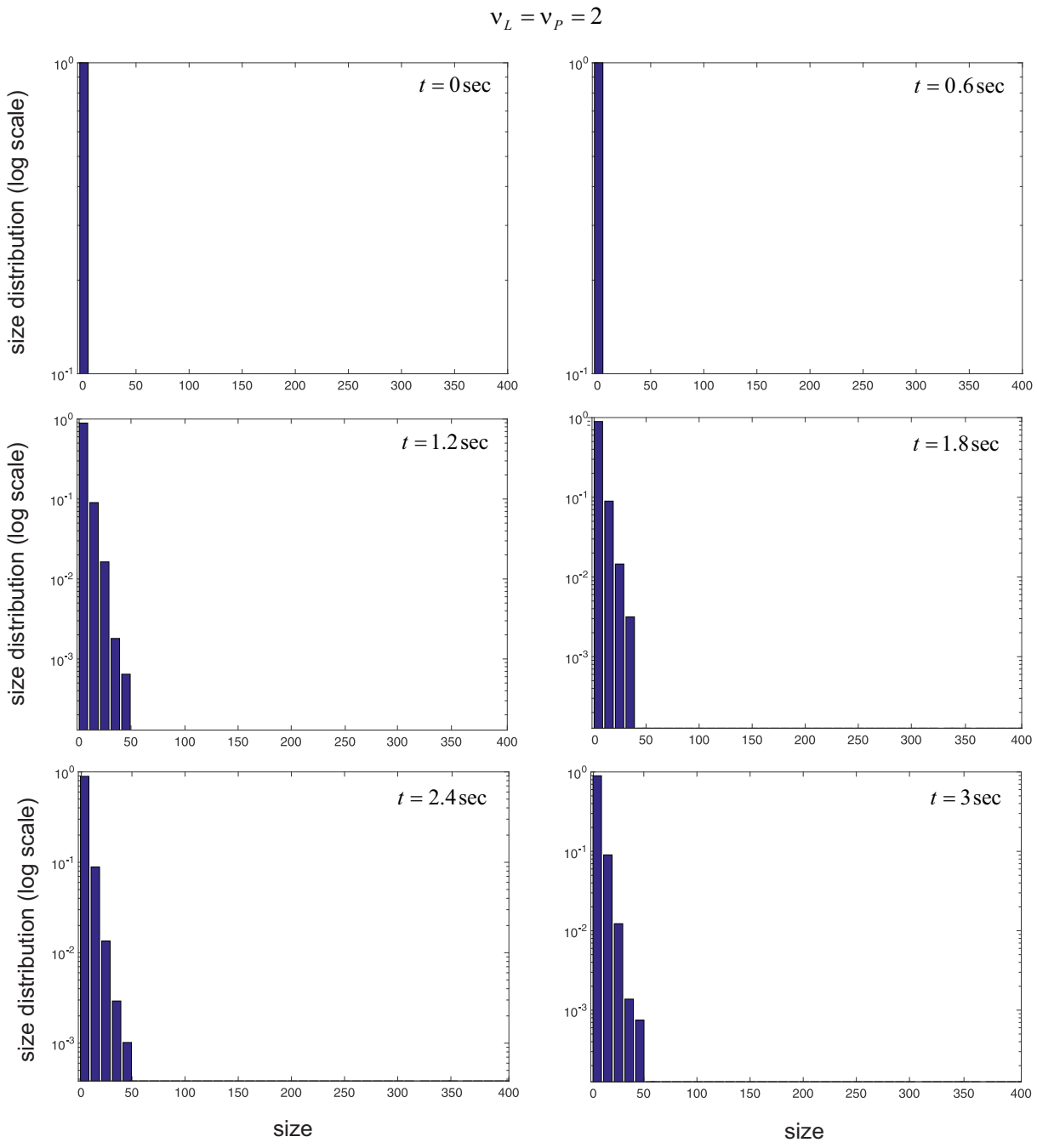

Figure 4.4: Estimated size distributions of molecular aggregates at different time points when $\nu_L = \nu_P = 2$.

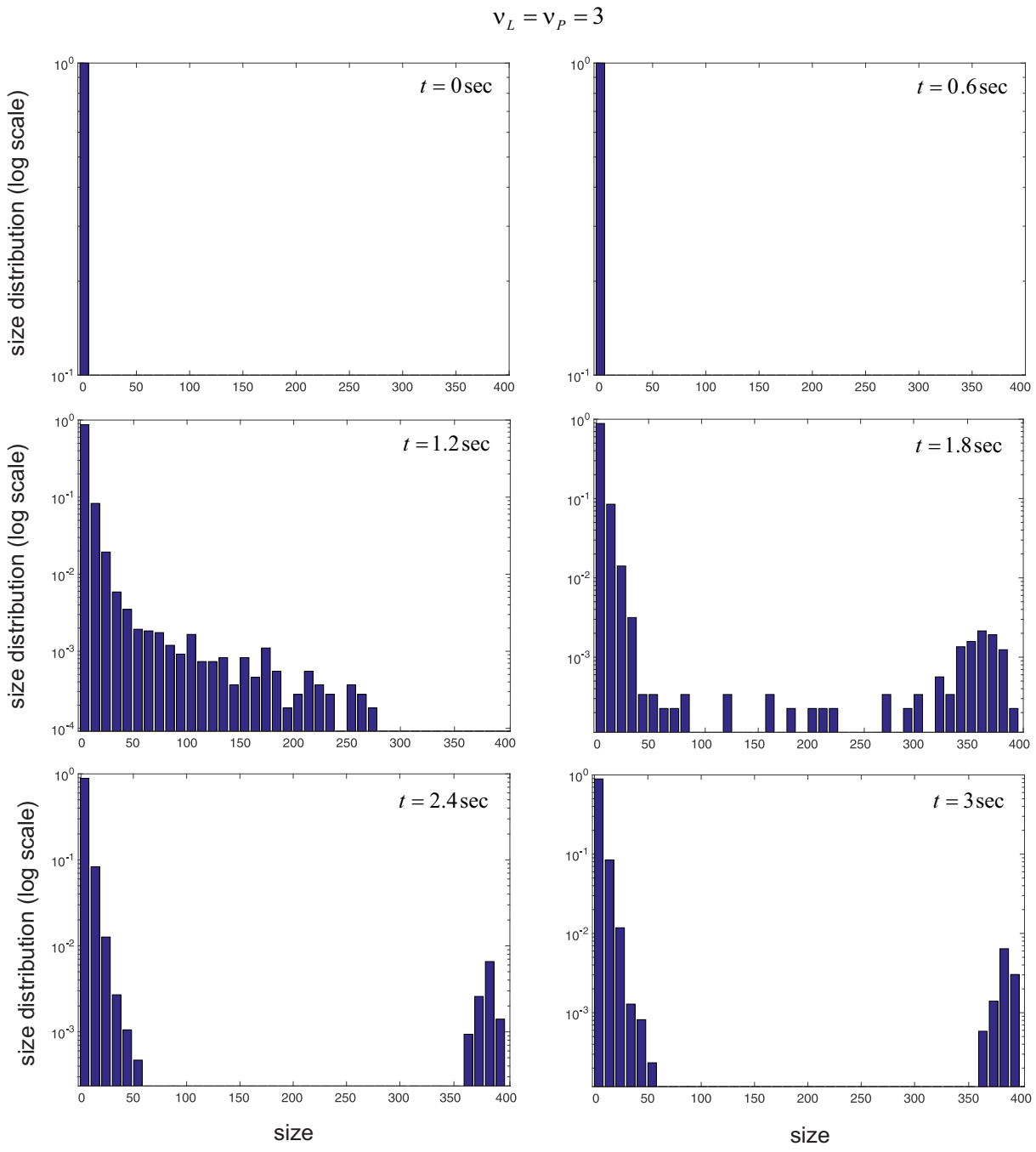

Figure 4.5: Estimated size distributions of molecular aggregates at different time points when $\nu_L = \nu_P = 3$.

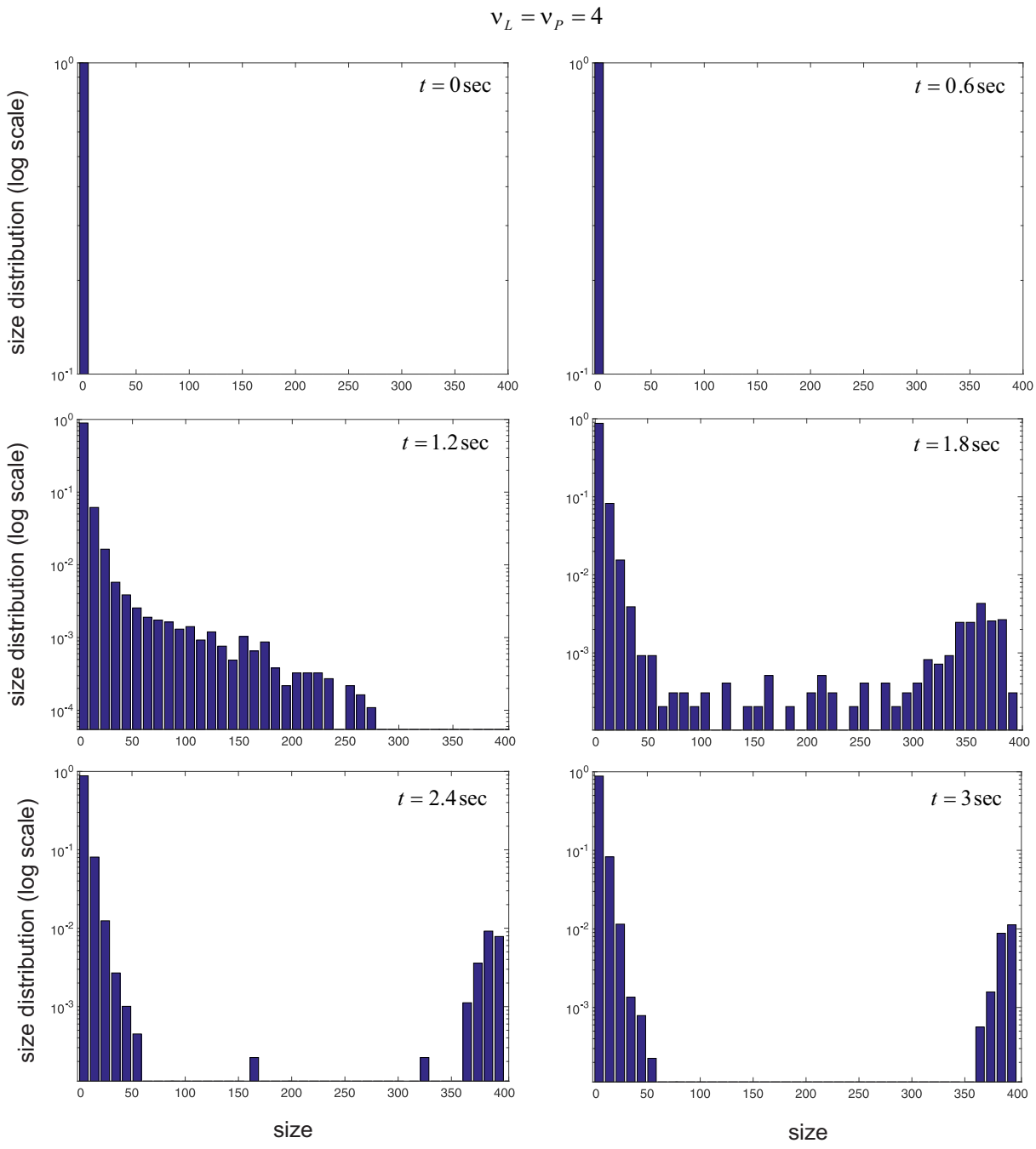

Figure 4.6: Estimated size distributions of molecular aggregates at different time points when $\nu_L = \nu_P = 4$.

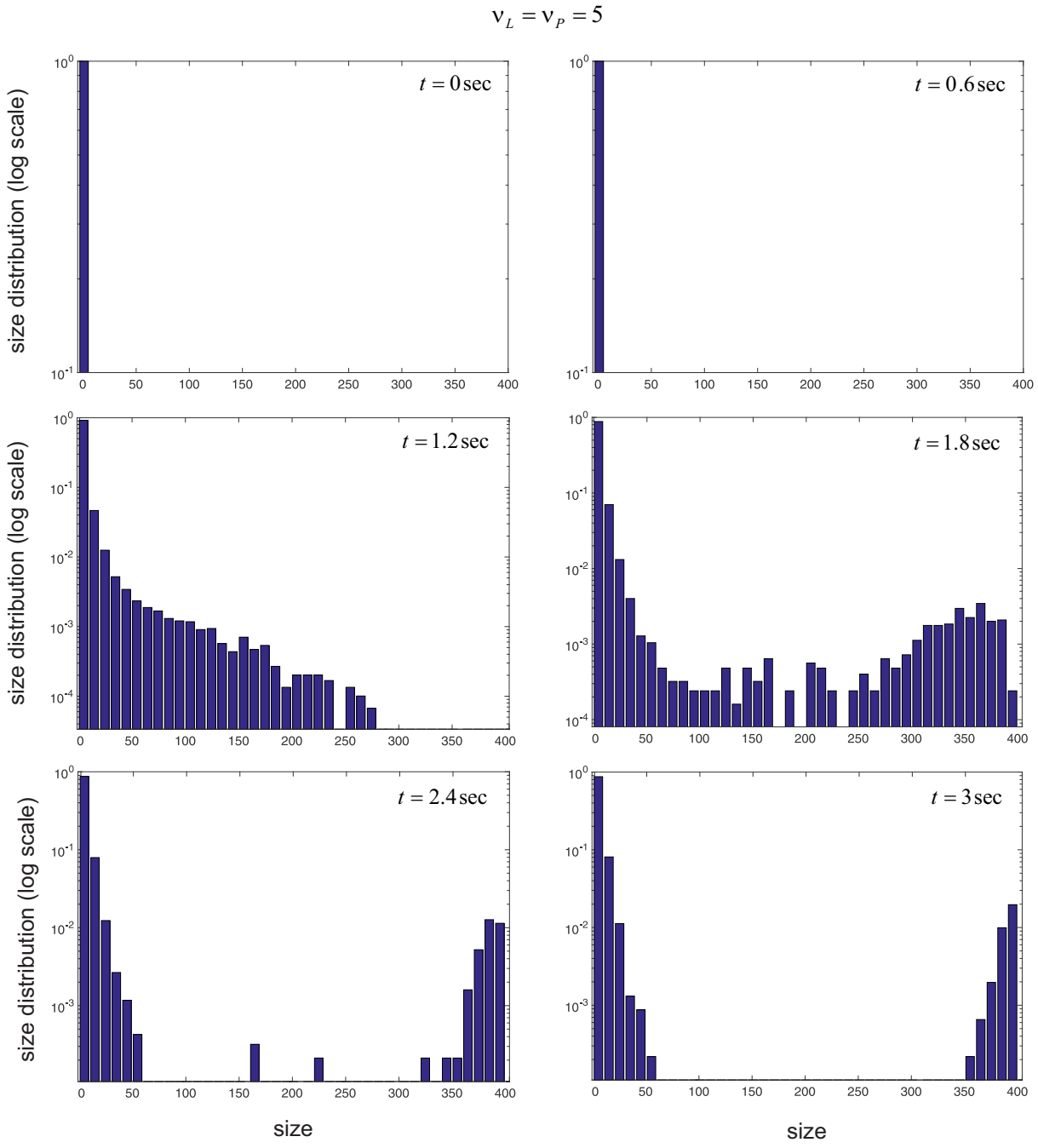

Figure 4.7: Estimated size distributions of molecular aggregates at different time points when $\nu_L = \nu_P = 5$.

formation of aggregates.

In Fig. 4.3, no formation of aggregates comprised of more than two basic molecules L and P takes place (the reaction system contains only a mixture of L, P and LDP molecules). Therefore, no formation of aggregates with size greater than 2 takes place at valence number 1. On the other hand, Fig. 4.4 indicates that no formation of aggregates with size greater than 50 is observed at valence number 2. This is expected, since a low valence number limits the combinatorial binding of molecules and thus the size of the resulting aggregates. However, as the valence number increases, more aggregates with larger sizes gradually form, highlighting the crucial role of valency in the formation of complex aggregates.

Of particular importance is the fact that, when $\nu_{_L} = \nu_{_P} \geq 3$, the RDME-based model evolves to a state characterized by the formation of a single large molecular aggregate at steady state ($t = 3sec$), which may coexist with simpler molecules of appreciably smaller sizes. This behavior demonstrates the feasibility of hydrogel-like network synthesis based on the stochastic formation of FKBP-rapamycin-FRB complexes. Note also that, when $\nu_{_L} = \nu_{_P} \geq 3$, the size distribution evolves to a bimodal distribution, indicating the occurrence of a sol-gel phase transition.

For our biochemical system shown in Fig. 4.8a, simulation results for different valencies of equivalent FKBP and FRB proteins are shown in Videos 1-5. The results indicate that individual unbound molecules at the beginning of each simulation are scattered uniformly within the volume. For higher valencies of three or greater, quick formation of

relatively large aggregates comprising many FKBP and FRB molecules occurred, while only small aggregates were seen in valence number two, and none observed for valence number one, as expected. Moreover, convergence to a stationary state was observed, characterized by the formation of a single hydrogel-like aggregate that may coexist with simpler and appreciably smaller molecules.

Further evidence of phase transition is demonstrated by the estimated probabilities of iPOLYMER to produce aggregates of a threshold size of 100 or larger for different valence numbers (Fig. 4.8b), as well as for different rapamycin concentrations (Fig. 4.8c). The sharp increase in the probability values observed in Fig. 4.8b indicates that efficient polymerization can be achieved when the individual valence numbers of FKBP and FRB are at least three, with the total valence number of FKBP and FRB molecules being at least six. On the other hand, the sharp decrease in the probability values depicted in Fig. 4.8c indicates that efficient polymerization requires a sufficient concentration of rapamycin. This implies that, in addition to the valence numbers of FKBP and FRB, the concentration of the dimerizing agent is expected to directly affect phase transition. Note that the base number of rapamycin molecules is scaled to the common valency of FKBP and FRB. For example, for the case of valency 1 for FKBP and FRB, we have five different systems that are represented by five data points on the plot. The first system initially contains 40 rapamycin molecules, a number that is calculated by 40 (base number of rapamycin molecules) multiplied with 1 (the valency of FKBP/FRB). The second system initially contains 80 rapamycin molecules, a number that is calculated by 80 (base number of rapamycin

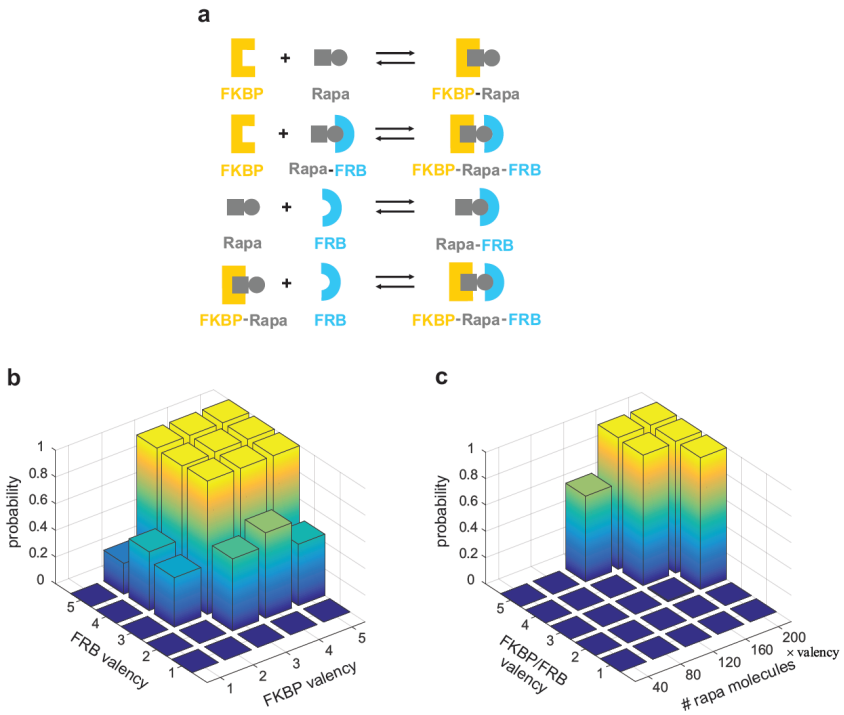

Figure 4.8: *In silico* implementation of iPOLYMER demonstrates its feasibility for hydrogel network synthesis. (a) Four reversible reactions between monomeric FKBP, FRB and rapamycin molecules modeled in our simulations. Each binding unit in the tandem repeats of FKBP or FRB can undergo the four reactions in the presence of rapamycin. (b) Estimated probabilities that iPOLYMER will produce aggregates of a threshold size of 100 or larger for different valence numbers of the FKBP and FRB molecules. An aggregate of size 100 comprises 25% of the total number of FKBP and FRB molecules initially present in the simulated system. (c) Estimated probabilities that iPOLYMER will produce aggregates of a threshold size of 100 or larger for different valence numbers of the FKBP and the FRB molecules, and different initial numbers of rapamycin molecules, determined by the base number of rapamycin molecules multiplied by their valency.

molecules) multiplied with 1 (the valency of FKBP/FRB), and so on.

To validate our computational results we obtained thus far, our collaborators performed experiments to evaluate iPOLYMER in living cells. Towards this end, they first generated two series of engineered proteins in order to track their expression in cells (Fig. 4.9a,b): a

yellow fluorescent protein (YFP) attached to N tandem repeats of an FKBP domain (denoted by YF_N, N = 1,2,3,4,5), and a cyan fluorescent protein (CFP) attached to similar tandem repeats of an FRB domain (CR_M, M = 1,2,3,4,5). Then, they first co-expressed the highest-valence number pair (YF_5 and CR_5), in COS-7 cells, in order to confirm diffuse fluorescence, added rapamycin at a relatively high concentration (333 nM), and imaged the fluorescence. Cells with high expression of both YF_5 and CR_5 peptides initially exhibited diffuse fluorescence signals that rapidly turned into puncta upon rapamycin addition (Fig. 4.10a), which steadily grew in size during prolonged rapamycin treatment. Finally, they carried out the experiment by acquiring Förster resonance energy transfer (FRET) measurements from CFP and YFP on the proteins, and observed increasing FRET values in the cytosol within 5 minutes of rapamycin addition.

FRET is a mechanism for capturing energy transfer between two light-sensitive molecules (i.e., chromophores) [54]. A donor chromophore, initially in its electronic excited state, may transfer energy to an acceptor chromophore through nonradiative dipoledipole coupling. FRET is extremely sensitive to small changes in distance since the efficiency of this energy transfer is inversely proportional to the sixth power of the distance between donor and acceptor. This makes FRET a powerful measuring technique for molecular interactions and bindings. For monitoring complex formation between two molecules, one of them is labeled by a donor and the other by an acceptor. In this way, the FRET efficiency is measured and used to identify interactions between the labeled molecules.

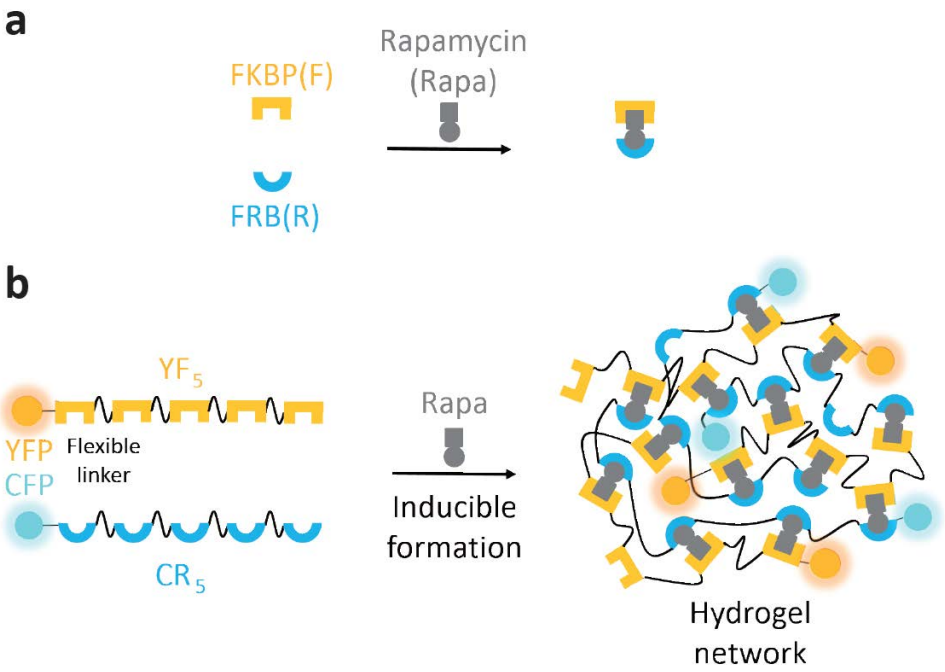

Figure 4.9: Schematic illustration of iPOLYMER. (a) Rapamycin induces rapid, stable and specific binding between FKBP and FRB molecules. (b) YF_5 and CR_5 contain five repeats of FKBP and FRB, respectively, spaced by 12 amino acid linker sequences. Mixing YF_5 and CR_5 (left) with rapamycin is expected to induce the formation of a hydrogel network (right). YF_N and CR_M contain N-repeats of FKBP and M-repeats of FRB with the same linkers, respectively.

In the experiment performed by our collaborators, a continuous FRET increase was observed at the puncta that emerged at later time points (see Fig. 4.10a). This implies that puncta formation was actually due to the binding between the two proteins induced by rapamycin, further supported by lack of these phenomena when dimethylsulfoxide (DMSO) was applied to the cells. DMSO is a solvent that has been shown to have no effect on a wide range of protein-protein interactions in cells [123].

Our collaborators also experimentally explored the effect of the valence numbers on the probability of puncta formation by testing all 25 different pairs of cytosolic proteins

cytoYF$_N$ and cytoCR$_M$, for N, M = 1,2,3,4,5, where cytoYF$_N$ and cytoCR$_M$ respectively contain N-repeats of FKBP and M-repeats of FRB. It is noteworthy that cytoYF$_N$ and cytoCR$_M$ essentially replicate our L and P molecules in the biochemical reaction system we defined in Fig. 4.1, and equivalently in Fig. 4.8a.

At small total valence numbers (i.e., when N+$M \leq 5$), less than 15% of cells formed puncta (Fig. 4.10b and Fig. 4.10c). By contrast, the percentage of cells with puncta increased rather dramatically at large total valence numbers. The similarity of the dependency of aggregate formation on valence numbers in living cells (Fig. 4.10b) to that *in silico* (Fig. 4.8b) strongly suggests that formation of puncta *in situ* is actually a result of FKBP/FRB polymer networks having undergone a phase transition. We could thus experimentally observe the dependency of both puncta formation kinetics and efficiency on valence number, which is in agreement with expectation drawn from our theoretical considerations.

In addition to the results discussed above in terms of the dynamics of the biochemical reaction system under consideration, since we now know that our computational model can generate molecular aggregates under certain conditions, we can use the model to generate graph representations corresponding to aggregates obtained by the RDME-based approach. Next, we discuss how we can construct such representations and use them to gain further insights into the sieving properties of molecular aggregates obtained in this work.

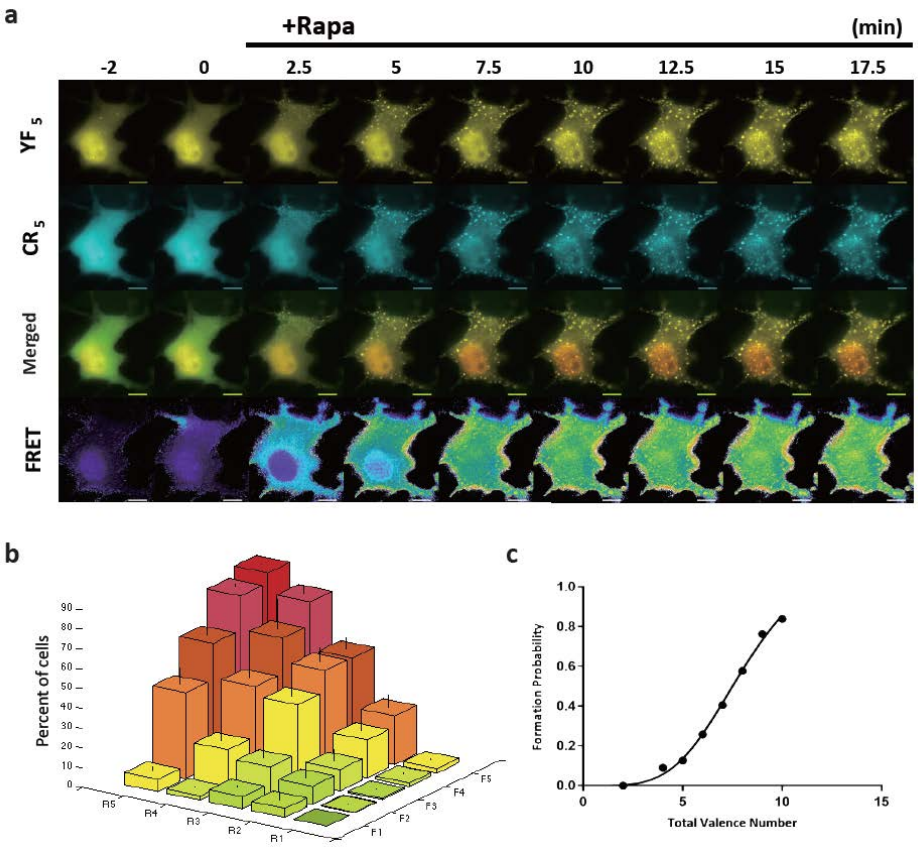

Figure 4.10: iPOLYMER puncta formation in living cells. (a) Time-lapse imaging of fluorescent puncta formation in COS-7 cells at indicated times relative to the addition of rapamycin. Scale bars, $10\mu m$. Punctate structures enriched with CFP, YFP, and FRET signals start to emerge within 5 min after rapamycin addition. (b) Frequency of iPOLYMER puncta formation plotted against valence numbers in FKBP and FRB constructs. F_N represents valence number of cytoYF$_N$, whereas R_M represents cytoCR$_M$. (c) Probability of iPOLYMER formation was plotted against the total valance number $N + M$. In order to avoid bias, combinations of (N, M) with either N or M being one were excluded from the data, except N=M=1. Note that the peptide with single valency should not lead to network formation, confirmed by the rare puncta formation in (b).

4.4 Molecular Aggregates as Sieves and their Effective Pore Sizes

Molecular aggregates, such as the ones produced in this work, are known to form a selective three-dimensional gel-like sieve that interferes only slightly with the motion of small molecules but appreciably slows down the motion of large molecules and blocks them from passing through the sieve [93, 166]. To study the sieving properties of the aggregates generated by our RDME-based model, we will need to construct graph representations of these molecular aggregates. In the following, we discuss how we can construct these graphs from the results obtained by the RDME-based algorithm.

4.4.1 Graph Representation of Molecular Aggregates

A graph is determined by its vertices and edges. To define the vertices of the graph corresponding to a molecular aggregate generated by our computational model, we assign a unique label to the binding sites associated with the aggregate as follows. We label each unbound binding site on an L molecule as a *Type I* vertex. We also label each unbound binding site on a P molecule as a *Type II* vertex. Finally, we label each L-binding site that is bound to a P-binding site through a D molecule as a *Type III* vertex. We consider an L-binding site as being "unbound" if it is not bound to a P-binding site through a D molecule; i.e., if it is free or it is bound by a D molecule that is not bound to a P-binding

site. Similarly, we consider a P-binding site as being "unbound" if it is not bound to an L-binding site through a D molecule; i.e., if it is free or it is bound by a D molecule that is not bound to an L-binding site. To define the edges of the graph, we link the binding sites on L and P molecules that are physically linked to each other.

The reason for grouping vertices into Type I, Type II and Type III is that the actual physical lengths of the L, P and D molecules are different (Fig. 4.2). This becomes relevant when calculating the pore sizes of molecular aggregates, which will become clear in the next subsection. In Fig. 4.11, we illustrate our graph construction scheme by considering two molecules L and two molecules P with the L molecules having a valency of 4 and the P molecules having a valency of 3. In Fig. 4.11a, we depict a simple aggregate molecule made up of these four molecules together with five D molecules. The corresponding graph is depicted in Fig. 4.11b. This graph is represented by four Type I vertices, two Type II vertices, and four Type III vertices as well, by a total of ten edges.

Early in the simulation, as well as in cells or *in vitro*, when the L, D and P molecules start to interact and bind to each other, increasingly larger molecular aggregates are beginning to form. In Fig. 4.12, we depict an example of a non-trivial graph corresponding to such a molecular aggregate, which was computationally generated by our RDME-based model at $t = 1.5$sec (a time that is close to the onset of phase transition). This aggregate comprises a total of 40 L and P molecules with valency 5.

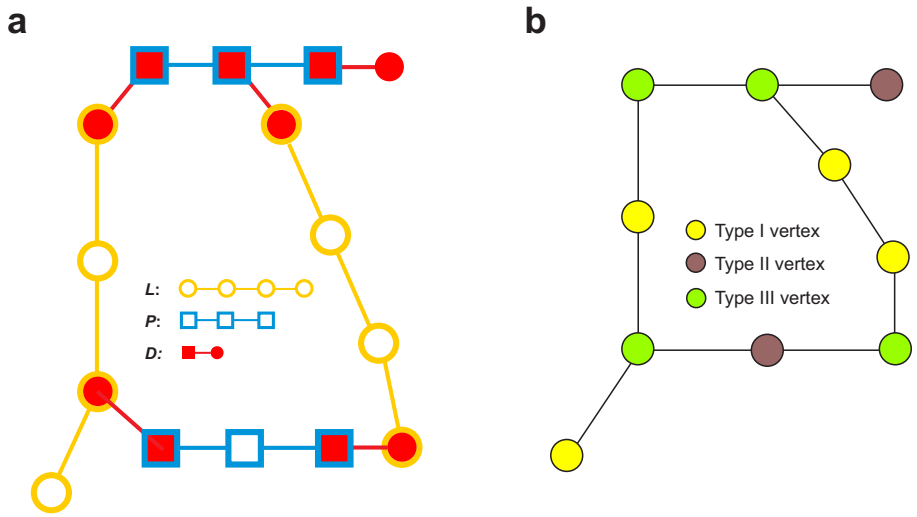

Figure 4.11: (a) An example of a simple aggregate molecule made up of two L molecules with valency 4, two P molecules with valency 3, and five D molecules. (b) The corresponding graph representation consisting of four Type I vertices, two Type II vertices, four Type III vertices, and ten edges.

4.4.2 Pore Size Distribution (PSD) and Effective Pore Size (EPS)

Now that our model can generate the graph representations corresponding to the molecular aggregates obtained by the RDME-based algorithm, to study the sieving property of a molecular aggregate of a given size, we can use its *pore size distribution* (PSD) at time t, defined by

$$\mathcal{P}_t(\sigma) = \frac{\mathrm{E}[P_t(\sigma)]}{\mathrm{E}[P_t]}, \quad \text{for } \sigma > 0,$$

where $P_t(\sigma)$ is the number of pores of size σ present in the aggregate, P_t is the *net* number of pores, and $\mathrm{E}[\cdot]$ denotes expectation. Unfortunately, we cannot calculate the PSD analytically. We can however approximate it computationally via Monte Carlo, provided that

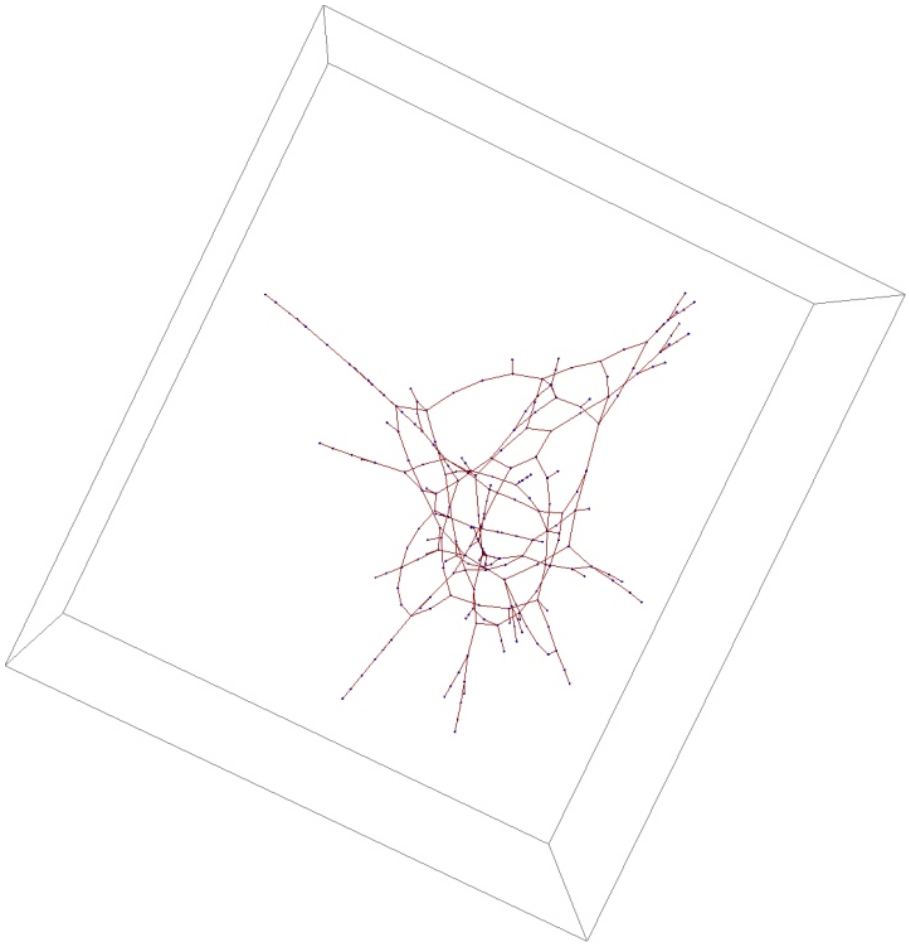

Figure 4.12: Graph representation of a molecular aggregate obtained by an RDME-based simulation at $t = 1.5$sec, which is close to the onset of phase transition. This aggregate comprises a total of 40 L and P molecules with valencies $\nu_{\scriptscriptstyle L} = \nu_{\scriptscriptstyle P} = 5$.

we can appropriately compute $P_t(\sigma)$. We can do this by kinetic Monte Carlo simulation runs of the RDME-based model that produces K instances of the aggregate at time t and by computing

$$\widehat{\mathcal{P}}_t(\sigma) = \frac{\sum_{k=1}^{K} P_t^{(k)}(\sigma)}{\sum_{k=1}^{K} \sum_{\sigma' \geq 1} P_t^{(k)}(\sigma')}, \tag{4.9}$$

where $P_t^{(k)}(\sigma)$ is the number of pores of size σ in the k-th aggregate instance obtained at time t.

An important use of the PSD is to determine which molecules can pass through a sieve and which molecules will be blocked. Let us denote by $\pi_t(\sigma)$ the (expected) probability that molecules of *physical size* σ pass through a sieve present at time t.[4] In this case, the (expected) probability that molecules of *physical size* σ are blocked by the sieve is given by $1 - \pi_t(\sigma)$. Note that

$$\pi_t(\sigma) = \sum_{\sigma' \geq \sigma} \mathcal{P}_t(\sigma').$$

We can therefore approximate the probability $\pi_t(\sigma)$ by computing

$$\widehat{\pi}_t(\sigma) = \sum_{\sigma' \geq \sigma} \widehat{\mathcal{P}}_t(\sigma') = \frac{\sum_{\sigma' \geq \sigma} \sum_{k=1}^{K} P_t^{(k)}(\sigma')}{\sum_{\sigma' \geq 1} \sum_{k=1}^{K} P_t^{(k)}(\sigma')}. \tag{4.10}$$

Let us now define the size σ_t^*, given by

$$\sigma_t^* = \max \Big\{ \sigma \mid \widehat{\pi}_t(\sigma) \geq 1/2 \Big\}. \tag{4.11}$$

This is the maximum size of a molecule that is expected to *most likely* pass through the sieve. Moreover, molecules with sizes $\sigma > \sigma_t^*$ will *most likely* be blocked by the sieve. We refer to σ_t^* as the *effective pore size* (EPS) of the sieve that is formed at time t.

Although we can potentially compute the EPS from the *pore size distribution* (PSD)

[4] We represent a molecule by the smallest containing sphere and define its *physical size* to be the length of the sphere's diameter.

of the molecular aggregate, we were unable to calculate the PSD or obtain the exact value of the EPS for the biomolecular sieves we constructed in this work due to experimental difficulties. However, our experimental collaborators were able to estimate an upper bound of the EPS for our experimentally synthesized gels, either in cells or *in vitro*, through a series of independent experiments that we discuss later in this chapter.

4.4.3 Estimation of PSD and EPS

In order to gain further insight into the sieving property of a gel, we now seek to estimate the EPS of molecular aggregates that are computationally generated by the RDME-based model. However, it is important to note that calculating PSDs for non-trivial dense aggregates, which are the types of aggregates obtained at steady state in systems in which phase transition takes place, is a computationally intractable problem. This is because calculating PSDs would require identifying the "pores" on graphs that correspond to these aggregates, which may comprise several hundreds or even thousands of vertices and edges. This increase in graph size and/or density would dramatically affect the computational complexity of the problem. On the other hand, due to the dense structure of the graphs, we can infer a plausible range for the EPS values of dense molecular aggregates by utilizing experimental measurements of the lengths of their vertices and edges, which could serve as a basis for assessing the experimentally obtained EPS bounds. We perform this analysis in the next section.

In addition, the problem of estimating the EPS value for a molecular aggregate formed

at an early time is experimentally very challenging and we were not able to perform such experiments either in cells or *in vitro*. Therefore, and in the context of a hypothetical scenario and through a concluding computational analysis, we seek to calculate the PSDs and EPSs for certain sets of molecular aggregates that are computationally generated before the system reaches steady state. We expect such aggregates to have relatively less dense graph structure with fewer vertices and edges than aggregates formed at a later time, which will make the identification of pores on these graphs computationally manageable. We carried out this analysis by employing a recently proposed graph-theoretic method [116], by identifying all existing pores in each graph instance of an aggregate, by estimating their sizes based on the actual physical length of the constituent molecules, as measured experimentally, and eventually, by calculating the PSD and EPS values for the aggregate. The following two types of analysis characterize a gel by its EPS value and provide further insight into the development of biomolecular sieves.

4.4.3.1 Estimating Pore Sizes of Molecular Aggregates at Steady State

In this section, we discuss the problem of inferring plausible estimates for PSDs of molecular aggregates at steady state. To do so, we first define pores as being one of two types of cycles in the 3D graph associated with an aggregate: (i) triangles, and (ii) chordless cycles. A cycle is a set of edges over a closed walk, which consists of a sequence of vertices starting and ending at the same vertex, with each two consecutive vertices in the sequence adjacent to each other in the graph. No repetitions of vertices is allowed, other

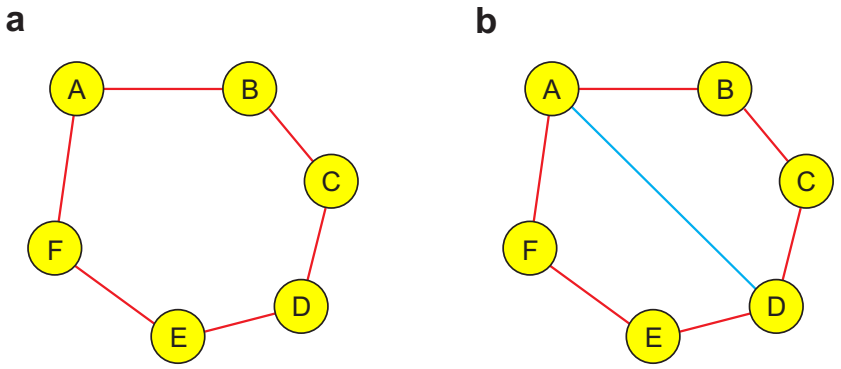

Figure 4.13: (a) ABCDEFA is a chordless cycle. (b) ABCDEFA is not a chordless cycle due to the presence of chord AD connecting vertex A to vertex D. However, the cycles ABCDA and ADEFA are chordless.

than the repetition of the starting and ending vertex. A triangle is a cycle of length three, whereas a chord is an edge connecting two non-consecutive vertices on a cycle. A cycle is said to be chordless if its length is at least four and has no chords (Fig. 4.13).

To estimate the size of a pore whose shape is polygon-like, we approximate the pore with a circle whose perimeter is taken to be the same as the physical perimeter of the cycle associated with the pore, and use the circle's diameter to quantify the size of the pore. We calculate the physical perimeter (girth) of a pore by adding the physical lengths corresponding to each type of vertex in the corresponding cycle as well as the physical lengths of its edges. According to Fig. 4.2, the physical length of a Type I vertex is 2.74 nm, of a Type II vertex is 2.20 nm, and of a Type III vertex is 5.00 nm. Note that we do not include "overhanging" D molecules (i.e., D molecules attached only to an L or a P molecule) when specifying the physical lengths of Type I and Type II vertices. These molecules are not part of the cycle and, therefore, are not relevant when calculating its physical perimeter. On the other hand, the physical length of an edge linking two consecutive binding sites on

an L or a P molecule has been experimentally determined to be equal to 6.00 nm. As a consequence, the size σ of a pore is given by the following simple formula

$$\sigma = \frac{1}{\pi}\left(n_{\mathrm{I}} \times 2.74 + n_{\mathrm{II}} \times 2.20 + n_{\mathrm{III}} \times 5.00 + n_e \times 6.00\right) \text{nm},$$

where n_{I} is the number of Type I vertices in the cycle corresponding to the pore, n_{II} is the number of Type II vertices, n_{III} is the number of Type III vertices, and n_e is the number of edges.

As an example, the graph depicted in Fig. 4.11b consists of only one chordless cycle corresponding to one pore. This cycle consists of $n_{\mathrm{I}} = 3$ Type I vertices, $n_{\mathrm{II}} = 1$ Type II vertex, $n_{\mathrm{III}} = 4$ Type III vertices, and $n_e = 8$ edges. As a consequence,

$$\sigma = \frac{1}{\pi}\left(3 \times 2.74 + 1 \times 2.20 + 4 \times 5.00 + 8 \times 6.00\right) \text{nm} = 24.97 \text{ nm}.$$

This would suggest that molecules of size (diameter) no more than about 25 nm can presumably pass through the pore.

During a simulation, as well as in cells or *in vitro*, increasingly larger molecules start to form and larger molecules also bind to each other and eventually form aggregates. As time proceeds, free binding sites of L and P molecules on the aggregate are likely to bind to each other. Thus, the aggregate will get denser with its pores becoming smaller. This is expected to occur in the presence of, or even without, new L and P molecules being added to the aggregate. In other words, the *cross-linking density* of the molecular aggregate, defined as

the ratio of the number of L-D-P complexes to the aggregate size (i.e., the number of L and P molecules on the aggregate), will presumably increase as a function of time.

At steady-state, the majority of the binding sites on the molecular aggregate will be bound. This means that it is very likely for two L and P molecules located next to each other to be bound. As a simple scenario, let us consider two adjacent L and P molecules on the aggregate, each with valency 5 (as was the case with our *in situ* and *in vitro* sieving experiments that determined the EPS). Given a sufficient number of available D molecules, the smallest pore that can be formed with the two L and P molecules corresponds to a triangle, based on the definition of the pore and the experimental measurements for the length of vertices and edges that was discussed earlier in this section. It turns out that, in our problem, triangles (i.e., cycles of length three) are characterized by three possible sizes: 9.62 nm, 9.79 nm, or 10.51 nm.[5] Since these values are very close to each other, we can assign an average size of $(9.62 \text{ nm} + 9.79 \text{ nm} + 10.51 \text{ nm})/3 = 9.97$ nm, or about 10 nm, to the pore size of any triangle on the graph of the aggregate. Moreover, the largest pore that the two L and P molecules can form corresponds to a chordless octagon with a size between 23.91 nm and 28.03 nm, an observation that is based on a similar analysis that depends on how the octagon is connected with the aggregate.[6] As one can imagine, the

[5]This is due to the fact that L and P molecules which form a triangle lead to a graph that has two Type III vertices and a third vertex that can be Type I, Type II, or Type III. In this case, the perimeter of a triangle will be equal to the sum of the lengths of the two Type III vertices (2×5.00 nm $= 10.00$ nm), of the length of the third vertex (which is 2.74 nm if Type I, 2.20 nm, if Type II, or 5.00 nm, if Type III, and of the length of the three edges (3×6.00 nm $= 18.00$ nm, since each edge is 6.00 nm long). Summing up the numbers associated with the three possibilities and dividing by π results in sizes 9.62 nm, 9.79 nm, and 10.51 nm.

[6]As a matter of fact, L and P molecules forming a chordless octagon that is connected to an aggregate lead to a graph that has at least three Type III vertices and at most eight Type III vertices. The octagon with three Type III vertices, together with two Type I and three Type II vertices, is the smallest octagon, whereas the octagon with eight Type III vertices is the largest octagon. In both cases, the length of the eight edges is given

two L and P molecules can alternatively form pores corresponding to a chordless square, pentagon, hexagon or heptagon with their pore sizes falling in the range between 9.97 nm (for a triangle) and 28.03 nm (for a chordless octagon).

We can extend the previous analysis to the case of more than two L and P molecules. In this case, the size of small pores will still fall within the range we obtained for the previous simple case of just two L and P molecules, whereas more L and P molecules could potentially form larger pores corresponding to polygons with more than eight edges. This would be particularly relevant for the aggregates that form earlier, since we expect these aggregates to have many unbound sites on their constituent L and P molecules, corresponding to a much lower cross-linking density, and be associated with relatively larger pore sizes. However, as we pointed out earlier in this section, an increasing number of binding sites on a molecular aggregate will become bound as the system reaches steady state. As a result, the aggregate will become increasingly denser, which would result in increasingly smaller pore sizes. Presumably, when the valence number of the constituent L and P molecules is 5 (consistent with our *in situ* and *in vitro* sieving experiments that determine the EPS), the sizes of these pores, and therefore the EPS value, will predominantly be within a range of about 10-28 nm, corresponding to polygons with relatively fewer number of edges.

From a practical perspective, and due to the 3D structure of an aggregate, it is likely that the spatial orientation of the molecules that form a dense aggregate could potentially block the passage of molecules through the pores and lead to a smaller EPS value. Taken

by 8×6.00 nm $= 48.00$ nm, since each edge is 6.00 nm long. Summing up the numbers associated with each of the two octagons and dividing by π results in sizes $(2\times2.74+3\times2.20+3\times5.00+8\times6.00)/\pi = 23.91$ nm and $(8 \times 5.00 + 8 \times 6.00)/\pi = 28.03$ nm.

together, the previous observations lead to the conclusion that the EPS values of dense molecular aggregates which are experimentally observed at steady state will be smaller than the computationally calculated EPS values, which are predicted to be within a range of about 10-28 nm. It is finally important to note that the density of an aggregate directly affects its EPS value and could be potentially influenced by: (i) the valence number of the constituent L and P molecules, (ii) the initial concentrations of the L, P and D molecules, and (iii) the experimental procedure used to obtain the aggregate. We will revisit these practical issues in the context of our experimental results in the Discussion section.

4.4.3.2 Estimating Pore Sizes of Molecular Aggregates at Early-stages

As noted in the previous section, for an early forming aggregate (e.g., at a time close to the onset of phase transition), its cross-linking density is expected to be much lower than that of a comparably sized aggregate observed at steady state. We therefore expect that an early-stage aggregate will be characterized by fewer and presumably larger pores and be associated with a relatively less dense graph structure (Fig. 4.12) for an example). We now seek to computationally characterize such an aggregate by its PSD and EPS in order to gain further insight into the sizes of its pores and investigate how these sizes are affected by the valence number of the constituent L and P molecules. To do so, we need to identify the pores of a given molecular aggregate, which implies that we must find all chordless cycles and triangles in the corresponding graph. We explain how we perform this analysis next.

To identify chordless cycles in a graph, we first need to compute the adjacency matrix

of the graph. This is a matrix that summarizes which vertices are adjacent to which other vertices in the graph. As a consequence, the adjacency matrix defines the graph. We can compute the adjacency matrix of the graph corresponding to a molecular aggregate by assigning unique labels to the vertices of the graph and by determining the connections between vertices by tracking the occurrence of binding and unbinding reactions during our kinetic Monte Carlo simulation.

The method we use to identify chordless cycles has been proposed by John Pfaltz [116]. In brief, for each vertex i of a graph, the method finds every chordless cycle starting from that vertex using a depth-first approach [168]. It begins with a vertex j that is adjacent to i, as determined by the adjacency matrix, and adds that vertex to a trial "cycle prefix," which stores all vertices that can potentially form a cycle. It subsequently identifies a vertex k that is adjacent to j and adds it to the cycle prefix as well. It then recursively iterates this process, with j taking the role of i and k taking the role of j, seeking to extend the cycle prefix through a new vertex l adjacent to k. If the new vertex turns out to be the original starting vertex, then a cycle has been found. If the new vertex is a different element of the cycle prefix, a new cycle has been found that is not through the starting vertex. In either case, the method checks to see if the cycle is chordless. If it is, then it stores the cycle information (represented by the set of unique labels assigned to its vertices), backs up one level in the recursion, tries some other vertex in the neighborhood of the last vertex in the cycle prefix, and continues with a new cycle search.

It is noteworthy that, since the method seeks to find all chordless cycles through all

vertices, the presence of tree-like structures in the graph would be detrimental to its performance – see Fig. 4.12 for an example of tree-like structures in a graph. Tree-like structures do not have any cycles in them. If they are not removed from the graph beforehand, the algorithm will be searching for cycles on them in a futile manner, which will result in a combinatorial increase of computational complexity. For this reason, the method first reduces the graph to its "irreducible spine" consisting of only chordless cycles and then proceeds with the cycle search [116].

Since our method finds only chordless cycles in a graph (whose length is at least four), it does not identify "triangles" (cycles with length three). As a result, we need to find a way to include the size of triangles when estimating the pore size distribution. In the previous subsection, we calculated the average pore size corresponding to a triangle on the graph of a molecular aggregate to be 9.97 nm. Unfortunately, identifying triangles individually in large graphs, such as the ones obtained by our approach, is a computationally intensive problem due to its combinatorial complexity. However, we can calculate the number n_t of all triangles in a given graph using a known result from graph theory [**?**], and include $n_t \times 9.97$ nm corresponding to all the triangular pores in our pore size calculation. It is shown in [**?**] that the number of all triangles in a graph is given by $n_t = \text{trace}[\mathbb{A}^3]/6$, where $\mathbb{A}$ is the adjacency matrix of the graph.

It turns out that using Pfaltz's method to identify pores in the graphs corresponding to aggregates with sizes above 40 (which are graphs obtained at times close to steady state) is computationally intensive and not practical with our current computational capabilities.

This is due to the fact such aggregates may produce graphs comprising several hundreds or even thousands of vertices. This dramatically increases the complexity of cycle search, making computation of the PSD intractable from a practical perspective.

To deal with this problem, we considered estimating the PSD of aggregates formed at early times away from steady state. Unfortunately, for aggregates with low probability of occurrence, Eq. (4.9) requires a large number of simulation runs, which can be computationally demanding. Note however that we expect molecular aggregates of comparable sizes to be structurally similar. For this reason, we included in Eq. (4.9) all aggregates of comparable sizes to the basic aggregate observed in a simulation run at time t, which appreciably reduced the number of simulation runs needed for approximately computing the PSD. We then focused on estimating the PSD of aggregates of sizes 30-40 formed at time $t = 1.5$ sec. We did so by independently simulating our RDME-based model $K = 2{,}000$ times for three different valencies: $\nu_L = \nu_P = 3, 4, 5$. For each simulation run, we identified all aggregates at time $t = 1.5$ sec with sizes 30-40 and computed the number $P^{(k)}_{t=1.5}(\sigma)$ of pores in these aggregates, for $\sigma > 0$. We then estimated the PSD using Eq. (4.9). The results are depicted in Fig. 4.14. By using Eqs. (4.10) & (4.11), we also estimated the corresponding EPSs $\sigma^*_{t=1.5}$. It turns out that $\sigma^*_{t=1.5} \simeq 60$ nm for valency 3, $\sigma^*_{t=1.5} \simeq 80$ nm for valency 4, and $\sigma^*_{t=1.5} \simeq 110$ nm for valency 5.

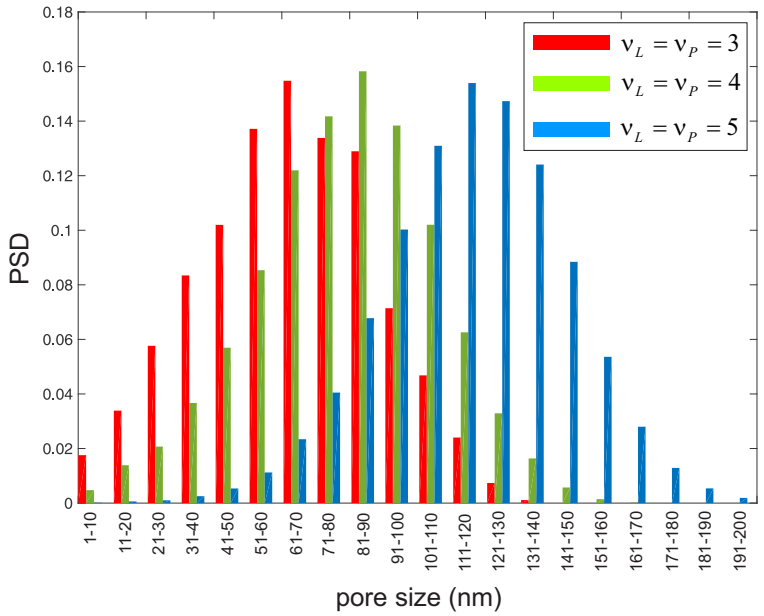

Figure 4.14: Estimated PSDs of molecular aggregates with comparable sizes of 30-40 nm, observed by our *in silico* implementation of iPOLYMER at time $t = 1.5$sec (close to the onset of phase transition). These distributions are binned into groups of 10 consecutive pore sizes. Clearly, polymerization of L and P molecules with larger valence numbers may result in early-stage aggregates with coarser sieving potential than molecular sieves formed by molecules with smaller valencies.

We would like to note here that the PSD of a molecular aggregate changes with time as aggregate formation evolves. It is reasonable to expect that, given a sufficient number of rapamycin molecules, the EPS of the sieve formed by a molecular aggregate at an earlier time will be larger than the EPS of the sieve formed by a molecular aggregate of comparable size at a later time. This is due to the fact that, as time proceeds, free binding sites of FKBP and FRB molecules that form the aggregate are likely to bind to each other, thus reducing the size of the pores. In other words, we expect that the cross-linking density will increase as a function of time, even without new FKBP and FRB molecules being added to the aggregate. As a matter of fact, aggregates comprising FKBP and FRB molecules with

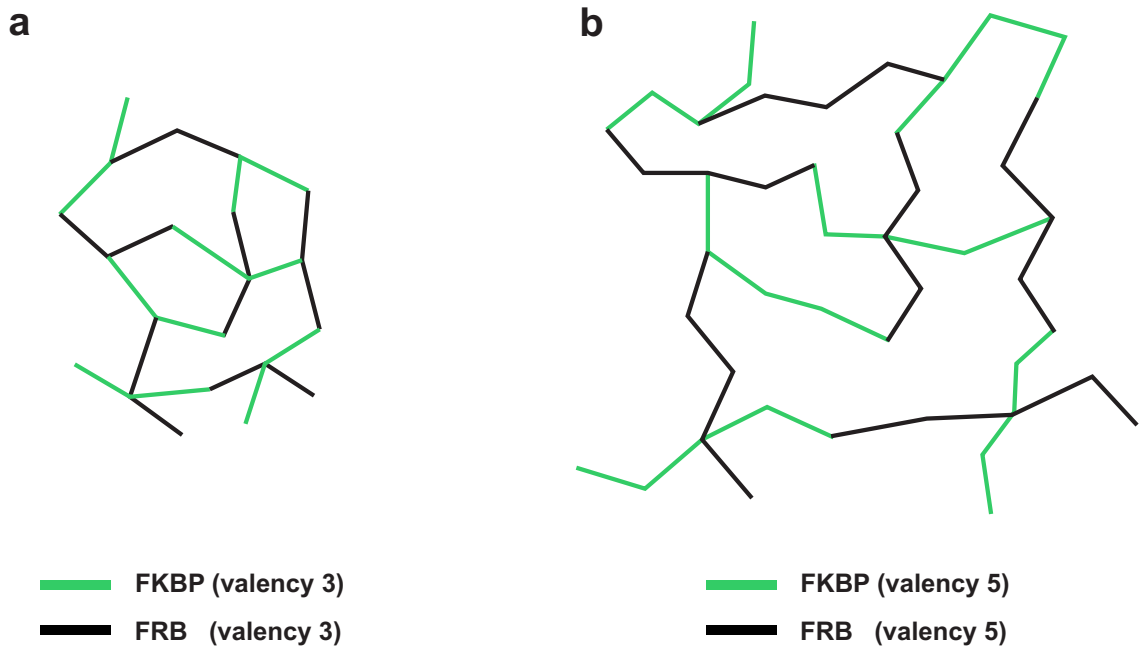

Figure 4.15: The aggregate in (a) is formed with FKBP and FRB molecules whose valencies are smaller than the valencies of the FKBP and FRB molecules forming the aggregate in (b). These aggregates have a relatively similar cross-linking pattern and identical cross-linking density of $5/4$, calculated by dividing the number (15) of FKBP-rapamycin-FRB complexes per each compound molecule with the number (12) of the FKBP and FRB molecules on each compound molecule. As a consequence, the aggregate in (a) is characterized by smaller pores than the aggregate in (b).

smaller valence numbers are originally formed by smaller constituent molecules (a smaller valence number corresponds to a smaller multivalent molecule in terms of its physical size), and these aggregates are more likely to be characterized by smaller pores earlier in their formation than aggregates comprising FKBP and FRB molecules with larger valence numbers (Fig. 4.15). This explains why, in the case of early-stage aggregates, the PSD depicted in Fig. 4.14 shifts to the right as the valence number of the constituent equivalent FKBP and FRB molecules increases.

The previously described method for identifying chordless cycles in a graph has been

implemented in a Unix platform using C++ by John Pfaltz [116]. It consists of two parts, which correspond to two executable files, GR_RED and CYCLE_DIST that are provided in [84].

The first executable file, GR_RED, accepts as an input the "edge list" representation of the graph, which lists all edges in the graph with their associated vertices. This list is obtained from the adjacency matrix of a given graph and is produced, in the form of a text file, by our MATLAB code that generates the graphs associated with the aggregates. GR_RED reduces a graph to its irreducible spine consisting of only chordless cycles. It also generates two output files, REDUCED and TRACE. The file REDUCED stores the information for the reduced graph to be used as input to the second executable file CYCLE_DIST. The file TRACE stores certain information on the progress of the code, which comes in handy when dealing with relatively large graphs. It is noteworthy that, for this type of graphs, the computational time required by the algorithm might be in the order of several hours to several days or even weeks, depending on the size and complexity of the graph. For this reason, we strongly recommend use of a computing cluster. In our work, we used a computing cluster consisting of 19 computing nodes, with each node being equipped with 24 processing cores and 128GB of memory (RAM).

The second executable file, CYCLE_DIST, reads the output file REDUCED generated by GR_RED. It also uses the text file PARAMETERS to read the values of the physical lengths (in nm) of the constituent L and P molecules and their valencies, together with the physical length of the L-D-P complex and the physical distance between consecutive

binding sites on L and P molecules. It then runs through the cycle search algorithm and identifies the chordless cycles represented by the sets of their vertices. It outputs the file CYCLE_LENGTHS, which stores all cycle lengths (pore sizes) in nm, as well as the lengths corresponding to triangles in the graph. This information is finally used to calculate the pore size distribution of the graph.

4.5 Discussion and Conclusions

In this research work, we developed a physical model for three-component multivalent-multivalent molecular interactions that led to a rigorous method for computationally implementing iPOLYMER. Our approach was based on a realistic kinetic Monte Carlo simulation algorithm that produced sufficiently accurate approximations of stochastic reaction-diffusion dynamics.

We noticed that aggregate formation occurred faster *in silico* than in cells or *in vitro* (See Figs. 4.5-4.7 and Fig. 4.10). We partially contributed this difference to the observation that our computational model does not take into account the fact that the rates of diffusion decrease as the sizes of the aggregate molecules become larger. Moreover, our model is based on the assumption that the rapamycin molecules are uniformly mixed with the FKBP and FRB peptides at the start of the simulation and, therefore, the model does not take into account the appreciable delay introduced initially by rapamycin diffusion into cells. Finally, our computational analysis was based on a system volume of $1(\mu\text{m})^3$, which

is much smaller than the actual cellular volume of about $25(\mu\text{m})^3$, resulting in faster convergence to steady state. Despite these differences, however, the results obtained by computational analysis provided valuable insights into the qualitative behavior of iPOLYMER, which could not be easily obtained experimentally.

Our RDME-based model validated aggregate synthesis for sufficiently high valence numbers of the constituent L and P molecules. In addition, the model captured the occurrence of phase transition, in the form of the size distribution evolving into a bimodal distribution, which indicated coexistence of large aggregate molecules with simpler molecules of appreciably smaller sizes (Figs. 4.5-4.7 and the accompanying Videos 3-5). It moreover demonstrated the fact that phase transition depends on the valence numbers of the L and P molecules and on the concentration of the dimerizing agent rapamycin (Fig. 4.9b,c). Our *in silico* results provided strong supporting evidence to our experimental results associated with hydrogel-like network synthesis (see Fig. 4.10a and our manuscript [84]), as well as with the dependency of network synthesis on the valence number of the constituent L and P molecules (Fig. 4.10b,c), and on the concentration of rapamycin (reported in [84]).

To investigate the sieving properties of molecular aggregates produced by iPOLYMER *in silico*, we estimated the pore sizes on the graphs corresponding to molecular aggregates generated by our RDME-based model. We did so for two different groups of non-trivial aggregates: (i) dense aggregates observed at steady state, and (ii) early-stage aggregates observed at a time close to the onset of phase transition. We considered these two groups separately, since calculating PSDs for non-trivial dense graphs, such as the ones generated

by our model at steady state, is computationally intractable.

Because of the previous unavoidable limitation of our model, we expect that the actual EPS values of the sieves will be smaller than the ones we estimated for dense aggregates at steady state and for early-stage aggregates obtained close to the onset of phase transition. This is because the spatial orientation of molecules that form an aggregate will physically block passage of molecules due to the dense structure of the aggregate (i.e., higher cross-linking density) and the presence of molecules in front of the pores.

The *in situ* and *in vitro* experiments yielded dense aggregates, taking into account that there were sufficient concentration levels of rapamycin in the system and the fact that the aggregates were observed at steady state. For this case, our computational analysis produced an EPS value estimate within a range of 10-28 nm. We expect that, since our computational model cannot consider the 3D spatial orientation of the constituent peptides on the aggregate, the actual EPS value would be lower.

The corresponding experimental results regarding the pore sizes of aggregates produced by our collaborators are discussed in our manuscript [84] in detail. In brief, from the *in vitro* experimental results, the EPS value was estimated to be within a range of 4.3-6 nm, based on the observation that fluorescent tracers (4.3 nm in diameter) did penetrate into hydrogels almost freely, whereas Q-dots (6 nm in diameter) and fluorescent beads (20 nm in diameter) did not. On the other hand, from the *in situ* experimental results, the EPS value was estimated to be within a range of 16-70 nm. This is supported by the evidence that mCherry-β-galactosidase, a tetramer complex with a rough diameter of 16 nm, passed

through the gel, whereas there was no single event that mCherry-TGN38, an mCherry-labeled vesicle with a varied size in the range 70-140 nm, could pass through the gel. Notably, the experimental predictions are in line with our computational prediction, which provided another level of validation of our results.

The estimated range of EPS values (4.3-6 nm) of the gel formed *in vitro* turned out to be considerably smaller than the one obtained *in situ* (16-70 nm). Presumably, this is due to the distinctly higher *in vitro* concentrations of the peptides and to the possibility that our EPS estimates have been affected by the fact that the gels obtained from the *in vitro* experiments were centrifuged in an effort to demonstrate their structural integrity and their ability to hold water [84].

Regarding the effect of the valence number on the EPS value of a dense aggregate, we should first note that we inferred the 10-28 nm range by assuming that the valence number of the L and P molecules is 5. We argued, however, that the actual EPS value could be smaller due to the spatial orientation of the constituent molecules, which cannot be captured by our graphical representation of the aggregate. For the same reason, we also expect that the actual EPS value will further decrease (but presumably not significantly) at higher valence numbers, due to an increasing cross-linking density and, consequentially, due to a more pronounced effect of the spatial orientation of the constituent molecules.

Our computational analysis, applied on early-stage aggregates, demonstrated a dependency of the EPS value on valency and rapamycin concentration (Fig. 4.9b,c and Fig. 4.15), although experiments could not be performed to validate this behavior. However, our col-

laborators have experimentally demonstrated that the relative concentration of rapamycin to the concentration of the L and P molecules could affect the density and even the formation of aggregates (see Fig. 4.9c, and [84] for a more detailed discussion).

It is noteworthy that, in a recent paper [89], Li *et al* proposed an algorithm to simulate stochastic two-component multivalent-multivalent interactions. The method involves the association/dissosiation of two types of molecules (proteins) with the binding domains of the first molecule type having affinity for the binding domains of the second molecule type. Although, at a first glance, this method seems to be similar to our method, there are some major and important differences. The most striking difference is that the Li *et al* method cannot be directly related to any physical model (such as the Doi model) for multivalent binding/unbinding interactions in continuous time/space. This method is based on a uniform discretization of time and, similarly to our method, on a uniform discretization of the three-dimensional space into voxels of equal volume. Formulas for the binding and unbinding probabilities are determined by using simple probabilistic arguments, and the same is true for molecular diffusion. It is not clear whether the method converges to a continuous time/space model as the time-step size and the voxel volume decrease towards zero. As a consequence, the Li *et al* method leads to an *ad hoc* algorithm that cannot be directly related to a physical model for multivalent molecular binding. This deficiency can seriously compromise the utility and accuracy of this method in an experimental setting.

On the other hand, our method models three-component multivalent-multivalent molecular interactions and provides a rigorous discretization of the well-known continuous

time/space Doi model for systems that involve reactions among different types of molecules as well as molecular diffusions. In an effort to guarantee that the resulting discretization converges to the Doi model, as the voxel volume approaches zero, our method provides appropriate formulas for the probability rates of the underlying binding/unbinding reactions as well as for the probability rates of molecular diffusion. Moreover, the proposed method treats time as a continuous variable and results in a realistic kinetic Monte Carlo simulation algorithm that is expected to produce sufficiently accurate approximations of stochastic reaction-diffusion dynamics.

Another fundamental disadvantage of the Li *et al* method is that it cannot provide a graphical representation of molecular aggregates. As a consequence, this method cannot be used to study sieving properties of molecular aggregates, e.g., by means of computing PSDs. The reason for this deficiency is that the method does not consider reactions among the binding sites of the underlying reactant molecules. Instead, it "coarsely" models aggregate formation by simply simulating stochastic binding and unbinding reactions between individual molecules using a rather *ad hoc* set of association and dissociation probabilities.

We should finally note that the technique developed in this dissertation is related to kinetic Monte Carlo methods for rule-based modeling of biochemical reaction systems [23, 138, 175]. Unlike conventional approaches that use population-based reactions to model biochemical reaction systems, these methods are based on rules which identify the molecular components involved in a transformative reaction and determine how these components change upon occurrence of a reaction under certain conditions. The main advantage of

rule-based methods is an appreciable reduction in required memory and computational cost when compared to conventional population-based approaches. This is particularly important for the problem at hand, since the number of distinct reactions that can occur may explode due to the formation of an increasing number of molecular aggregates. Moreover, rule-based approaches are capable of tracking the states of individual molecules while conventional approaches usually track only populations of chemical species.

Chapter 5

Conclusion and Future Directions

The exquisite orchestration of molecular interactions in cells is essential for the normal homeostatic regulation of multicellular organisms. Systematic delineation of networks of such molecular interactions is a challenging task. Moreover, the identification of interaction networks dysregulated in a particular disease may have profound effects on understanding the molecular causes that lead to the disease and may dramatically influence the development of effective strategies for pharmaceutical and therapeutic intervention.

In this research work, we introduced IntegraMiR, a novel computational method for inferring dysregulated miRNA/TF-mediated regulatory loops and networks that appear in a statistically over-represented manner in gene regulatory networks. IntegraMiR addresses the problem of miRNA-target prediction by appropriately constraining the statistical analysis of given mRNA/miRNA expression data and sequence-based target identification methods using relevant motif structures built by "prior" biological information readily available

in existing databases. The main strength of IntegraMiR originates from its capacity to fuse information from multiple sources and incorporate several statistical techniques to exploit almost any accessible aspect of available information in the expression data to identify integrated regulatory loops and networks at the transcriptional, post-transcriptional and signaling levels. Therefore, IntegraMiR adds to the ongoing effort of developing effective computational techniques for network inference by utilizing available experimental data and existing biological knowledge in an effort to produce reliable predictions in a context-dependent manner. With regards to the problem of network inference, we considered certain biological settings and relevant experimental data in the context of prostate cancer, as well as of Autism Spectrum Disorders (ASDs). It is noteworthy that the potential future directions we discuss in the following are relevant to the findings regarding both of these biological contexts.

To appropriately constrain the problem of predicting miRNA-target interactions, IntegraMiR focuses on specific types of three-node regulatory motifs: FFLs and Type III loops, both of which have attracted a great deal of attention in the literature. By identifying instances of dysregulated FFL and Type III motifs, and by using these motifs to construct interaction networks, IntegraMiR can also provide instances of two types of dysregulated two-node motifs: miRNA-TF negative and double-negative feedback loops.

The two key hypotheses behind our interest in Type III loop motifs are that miRNAs play major roles in regulating signaling pathways due to their sharp dose-sensitive nature, and that targets of single miRNAs are more connected (i.e., interact) at the protein level

than expected by chance. IntegraMiR identifies closely related miRNA targets on pathways deemed to be important in prostate cancer and delineates certain miRNA-mediated three-node regulatory loops in the KEGG Prostate Cancer Pathway. Some of the resulting network structures obtained from Type III loops represent dense overlapping regulon (DOR) motif [3] in which several input miRNAs co-regulate a set of output genes (known as a regulon). Co-targeting in a DOR pattern presumably strengthens the notion that the miRNAs involved share similar regulatory roles. The three-node loop motifs considered in this work can serve as basic building blocks for identifying more complex regulatory motifs, such as Single Input Modules (SIMs) and DORs [2, 26].

We should note that, in this work, we didn't address the problem of systematically constructing complex networks from the basic modules that are obtained by our procedure, which can be an exciting future direction of our research. Although our algorithm is capable of listing all three-node loops with nodes of interest, constructing a network from these three-node basic modules is an independent problem that has its mathematical roots in graph theory and graphical representations. Another potential future direction would be to directly incorporate in the algorithm the identification of four- or higher-node motifs, such as SIMs and DORs, in order to infer more complex network structures. However, one could still consider the problem of constructing complex networks from such higher-level motifs in such a scenario.

In principle, discoveries obtained by integrative computational approaches, similar to IntegraMiR, can provide systemic insights into the molecular biology of miRNA-mediated

interactions and can, thereby, assign context-dependent biological functions to poorly understood roles of miRNAs. With further advances in genomics research, the need for integrative analysis approaches capable of utilizing information acquired from various sources is becoming more evident than ever before. It is through these findings that researchers can form hypotheses aimed at accurately dissecting context-dependent molecular mechanisms underlying physiological and pathological conditions of interest. Through these types of analyses, effective drug targeting and successful disease treatments will eventually be realized. MiRNAs pose promising potential in this context.

IntegraMiR uses information from four databases, mSigDB, miRTarBase, TRANSFAC and TransmiR. If new and more informative databases become available in the future, information relevant to the problem discussed in this dissertation can be easily incorporated as part of the overall underlying strategy. For example, with the emergence and ever-increasing accessibility of high-resolution transcriptome data, by means of chromatin immunoprecipitation with sequencing (ChIP-Seq) experiments, together with regulation information, IntegraMiR could efficiently exploit such large-scale transcription factor-target information to obtain systems-level regulatory loops that could possibly account for much higher percentages in transcriptome changes.

We should note that a relatively large number of TF-target interactions are not included in the input to IntegraMiR owing to their unknown regulation type status in TRANSFAC and TransmiR. On the other hand, the method proposed in [174] does not utilize information on regulation type. As a result, although this method employs all available TF-target

interactions/associations, it cannot be used to identify coherent/incoherent FFL subtypes, which is *the* information required to derive a systems-level understanding of regulatory networks. However, by using all available TF-target interactions regardless of their regulation type and by limiting analysis to Type II FFLs, it was found in [174] that more than 20% of transcriptome changes could be attributed to these FFLs. This result demonstrates that FFL-based analysis has the potential to explain a considerable percentage of transcriptome changes. Once additional information about regulation type is made available through future database updates, we expect that IntegraMiR will produce results that are capable of explaining a higher percentage of transcriptome changes, with systemic insights similar to the ones presented in this work, as opposed to the approach in [174].

In constructing FFLs, IntegraMiR considers loops comprising miRNA and TF nodes that are both significantly dysregulated. The main reason for this choice is to focus primarily on FFLs that exhibit significant levels of dysregulation at both regulator nodes, which could play a major role in explaining observed transcriptome changes. This is mainly because our confidence that an FFL contributes to transcriptome dysregulation in prostate cancer will be diminished if the upstream regulator is differentially expressed but the downstream regulator is not (or vice versa). Note that IntegraMiR can be easily adjusted to identify FFLs in which at least one regulator node is significantly dysregulated. It is important however to understand that this adjustment, in combination with the high false-positive rate of sequence-based miRNA-target predictions, can result in an excessive number of predicted FFLs and relatively higher false-positive rates. This is due to the fact that this simple

modification allows a combinatorially larger number of potential nodes to be considered by the method.

Finally, the ranking score obtained by employing Fisher's method could be improved by using methods proposed to combine dependent statistical tests [14, 77]. However, due to lack of reliable between-node (and cross-platform) correlation estimation, accounting for dependencies is not feasible. Therefore, IntegraMiR uses Fisher's method to indicate the significance for each FFL by a ranking score, rather than a P-value. Upon availability of mRNA and miRNA expression data and techniques that could allow for reliable calculation of correlations, a possible future direction would be to incorporate such information into various aspects of the statistical analysis framework currently used by IntegraMiR to score FFLs more accurately.

In our final research work on modeling the dynamics of biomolecular interaction networks, we developed a physical model for three-component multivalent-multivalent molecular interactions that led to a rigorous method for computationally implementing iPOLYMER, a novel strategy developed by our collaborators at the School of Medicine to generate intracellular hydrogels that can act as biomolecular sieves. Our approach was based on a realistic kinetic Monte Carlo simulation algorithm that produced sufficiently accurate approximations of stochastic reaction-diffusion dynamics.

We should note here that several distinct models have been proposed in the literature to study stochastic and spatial effects in biochemical reaction systems (e.g., see [31, 46, 148]) and this is because no single model is currently capable of efficiently coping with the

broad range of spatial, temporal and concentration scales commonly found in biochemical reaction networks. For this reason, models such as the one discussed in this research, may represent a plausible approach that yields a compromise between computational efficiency as well as spatial and stochastic accuracy.

Our RDME-based model validates aggregate synthesis for sufficiently high valence numbers of the constituent L and P molecules. In addition, the model captures the occurrence of phase transition, in the form of the size distribution evolving into a bimodal distribution, which indicates coexistence of large aggregate molecules with simpler molecules of appreciably smaller sizes. It moreover demonstrates the fact that phase transition depends on the valence numbers of the L and P molecules and on the concentration of the dimerizing agent rapamycin. Our *in silico* results provided strong supporting evidence to the experimental results on hydrogel-like network synthesis that were obtained by our collaborators.

To investigate the sieving properties of molecular aggregates produced by iPOLYMER *in silico*, we estimated the pore sizes on the graphs corresponding to molecular aggregates generated by our RDME-based model. We did so for two different groups of non-trivial aggregates: (i) dense aggregates observed at steady state, and (ii) early-stage aggregates observed at a time close to the onset of phase transition. We considered these two groups separately, since calculating PSDs for non-trivial dense graphs, such as the ones generated by our model at steady state, is computationally intractable.

We therefore sought to infer a plausible estimate for the range of EPS values for these

aggregates based on certain assumptions and constraints of the problem without having to calculate the PSDs for such dense graphs. However, for early-stage aggregates, we were able to directly calculate the PSDs and approximately determine the corresponding EPS values. We considered molecular aggregates formed at a time close to the onset of phase transition. We expected that these aggregates would have a relatively lower cross-linking density, as compared to the ones obtained at steady state, and that the computationally estimated PSD values could quantify their sieving properties reasonably well.

We conjecture that, when a group of molecular aggregates with comparable sizes is characterized by a relatively high cross-linking density, the computationally estimated EPS values may overestimate the actual EPS values. This is due to the fact that our computational EPS estimation depends entirely on the graph structure of a molecular aggregate and the physical lengths of its constituent molecules, without taking into account the spatial orientation of the graph nodes. This is true because our model does not consider the actual distance between two nodes that are close but not adjacent to each other. It turns out that realistic modeling of the spatial orientation of constituent molecules in a molecular aggregate is an extremely challenging task, since it requires modeling of all dominant biochemical and biophysical energies and forces which give rise to a specific spatial configuration of molecules that form an aggregate.

A subsequent potential and at the same time challenging research problem would be to look into other graph-theoretic approaches developed by applied mathematicians and investigate their implementation feasibility to deal with large-scale dense graphs, such as

the ones produced by our RDME-based algorithm, to more accurately estimate their PSDs and consequently their corresponding EPS values.

In the end, we would like to note that the three specific research problems we tackled here, categorized under the two broad modeling paradigms in systems biology (i.e., network inference and modeling the dynamics of networks), have been identified as important applied research problems by our collaborators at the Johns Hopkins School of Medicine. The tools and techniques developed here have the potential to unravel the structure and the dynamics of certain complex intracellular interactions at a systems level, and can be generalized to many other biological settings. In this way, the findings from all three research problems present strong computational evidence that proves to be highly valuable to experimental biologists in providing reliable predictions and meaningful insights to help them guide their applied research in an efficient and cost-effective manner.

Bibliography

[1] A. S. Afshar, J. Xu, and J. Goutsias. Integrative identification of deregulated miRNA/TF-mediated gene regulatory loops and networks in prostate cancer. *PloS One*, 9(6):e100806, 2014.

[2] U. Alon. *An Introduction to Systems Biology. Design Principles of Biological Circuits*. Chapman & Hall/CRC, Boca Raton, FL, 2006.

[3] U. Alon. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, 8(6):450–461, 2007.

[4] S. Ambs, R. L. Prueitt, M. Yi, R. S. Hudson, T. M. Howe, F. Petrocca, T. A. Wallace, C.-G. Liu, S. Volinia, G. A. Calin, H. G. Yfantis, R. M. Stephens, and C. M. Croce. Genomic profiling of microRNA and messenger RNA reveals deregulated microRNA expression in prostate cancer. *Cancer Res.*, 68:6162–6170, 2008.

[5] S. Artmann, K. Jung, A. Bleckmann, and T. Beibarth. Detection of simultaneous group effects in microRNA expression and related target gene sets. *PLoS One*, 7(6):e38365, 2012.

[6] D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, 2008.

[7] L. A. Banaszynski, C. W. Liu, and T. J. Wandless. Characterization of the FKBP-rapamycin-FRB ternary complex. *J. Am. Chem. Soc.*, 127:4715–4721, 2005.

[8] D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.

[9] L. E. Becker, Z. Lu, W. Chen, W. Xiong, M. Kong, and Y. Li. A systematic screen reveals microRNA clusters that significantly regulate four major signaling pathways. *PLoS One*, 7(11):e48474, 2012.

[10] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.*, 57(1):289–300, 1995.

[11] L. Boldrup, P. J. Coates, M Wahlgren, G. Laurell, and K. Nylander. Subsite-based alterations in miR-21, miR-125b, and miR-203 in squamous cell carcinoma of the oral cavity and correlation to important target proteins. *J. Carcinog.*, 11:18, 2012.

[12] C. P. Brangwynne, T. J. Mitchison, and A. A. Hyman. Active liquid-like behavior of nucleoli determines their size and shape in xenopus laevis oocytes. *Proc. Natl. Acad. Sci. USA*, 108:4334–9, 2011.

[13] J. C. Brase, M. Johannes, H. Mannsperger, M. Fälth, J. Metzger, L. A. Kacprzyk, T. Andrasiuk, S. Gade, M. Meister, H. Sirma, G. Sauter, R. Simon, T. Schlomm,

T. Beibarth, U. Korf, R. Kuner, and H. Sültmann. *TMPRSS2-ERG*-specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-β signaling. *BMC Cancer*, 11:507, 2011.

[14] M. B. Brown. A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31:987–992, 1975.

[15] S. Chakraborty, S. Datta, and S. Datta. Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics*, 28(6):799–806, 2012.

[16] S. Chakraborty, S. Datta, and S. Datta. svapls: an R package to correct for hidden factors of variability in gene expression studies. *BMC Bioinformatics*, 14:236, 2013.

[17] L. W. Chang, A. Viader, N. Varghese, J. E. Payton, J. Milbrandt, and R. Nagarajan. An integrated approach to characterize transcription factor and microRNA regulatory networks involved in Schwann cell response to peripheral nerve injury. *BMC Genomics*, 14(1):84, 2013.

[18] A. Chaux, R. Albadine, A. Toubaji, J. Hicks, A. Meeker, E. A. Platz, A. M. De Marzo, and G. J. Netto. Immunohistochemistry for ERG expression as a surrogate for TMPRSS2-ERG fusion detection in prostatic adenocarcinomas. *Am. J. Surg. Pathol.*, 35(7):1014–1020, 2011.

[19] C. Y. Chen, S. T. Chen, C. S. Fuh, H. F. Juan, and H. C. Huang. Coregulation of

transcription factors and microRNAs in human transcriptional regulatory network. *BMC Bioinformatics*, 12(Suppl. 1):S41, 2011.

[20] K. Chen and N. Rajewsky. The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.*, 8(2):93–103, 2007.

[21] C. Cheng, K. K. Yan, W. Hwang, J. Qian, N. Bhardwaj, J. Rozowsky, Z. J. Lu, W. Niu, P. Alves, M. Kato, M. Snyder, and M. Gerstein. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput. Biol.*, 7(11):e1002190, 2011.

[22] B. Chopard, A. Dupuis, A. Masselot, and P. Luthi. Cellular automata and lattice Boltzmann techniques: An approach to model and simulate complex systems. *Advs. Complex Syst.*, 5:103–246, 2002.

[23] L. A. Chylek, L. A. Harris, C. S. Tung, J. R. Faeder, C. F. Lopez, and W. S. Hlavacek. Rule-based modeling: a computational approach for studying biomolecular site dynamics in cell signaling systems. *WIREs Syst. Biol. Med.*, 6:13–36, 2014.

[24] S. J. Cooper, H. Zou, S. N. LeGrand, L. A. Marlow, C. A. von Roemeling, D. C. Radisky, K. J. Wu, N. Hempel, V. Margulis, H. W. Tun, G. C. Blobe, C. G. Wood, and J. A. Copland. Loss of type III transforming growth factor-beta receptor expression is due to methylation silencing of the transcription factor GATA3 in renal cell carcinoma. *Oncogene*, 29(10):2905–2915, 2010.

[25] J. A. Copland, B. A. Luxon, L. Ajani, T. Maity, E. Campagnaro, H. Guo, S. N. LeGrand, P. Tamboli, and C. G. Wood. Genomic profiling identifies alterations in TGFbeta signaling through loss of TGFbeta receptor expression in human renal cell carcinogenesis and progression. *Oncogene*, 22(39):8053–8062, 2003.

[26] Q. Cui, Z. Yu, E. O. Purisima, and E. Wang. Principles of microRNA regulation of a human cellular signaling network. *Mol. Syst. Biol.*, 2:46, 2006.

[27] J. C. Darnell, S. J. Van Driesche, C. Zhang, K. Y. S. Hung, A. Mele, C. E. Fraser, E. F. Stone, C. Chen, J. J. Fak, W. Chi, D. D. Licatalosi, J. D. Richter, and R. B. Darnell. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, 146(2):247–261, 2011.

[28] B. De Craene and G. Berx. Regulatory networks defining EMT during cancer initiation and progression. *Nature Rev. Cancer*, 13:97–110, 2013.

[29] F. Demichelis and G. Attard. A step toward functionally characterized prostate cancer molecular subtypes. *Nat. Med.*, 19(8):966–967, 2013.

[30] A. Dhasarathy, D. Phadke, D. Mav, R. R. Shah, and P. A. Wade. The transcription factors Snail and Slug activate the transforming growth factor-beta signaling pathway in breast cancer. *PLoS One*, 6(10):e26514, 2011.

[31] M. Dobrzyński, J. V. Rodrguez, J. A. Kaandorp, and J. G. Blom. Computational

methods for diffusion-influenced biochemical reactions. *Bioinformatics*, 27:1969–1977, 2007.

[32] M. Doi. Second quantization representation for classical many-particle system. *J. Phys. A: Math. Gen.*, 9:1465–1477, 1976.

[33] M. Doi. Stochastic theory of diffusion-controlled reaction. *J. Phys. A: Math. Gen.*, 9:1479–1495, 1976.

[34] M. Dong, T. How, K. C. Kirkbride, K. J. Gordon, J. D. Lee, N. Hempel, P. Kelly, B. J. Moeller, J. R. Marks, and G. C. Blobe. The type III TGF-beta receptor suppresses breast cancer progression. *J. Clin. Invest.*, 117(1):206–217, 2007.

[35] Q. Dong, P. Meng, T. Wang, W. Qin, W. Qin, F. Wang, J. Yuan, Z. Chen, A. Yang, and H. Wang. MicroRNA let-7a inhibits proliferation of human prostate cancer cells *in vitro* and *in vivo* by targeting E2F2 and CCND2. *PLoS One*, 5(4):e10147, 2010.

[36] R. L. Elliott and G. C. Blobe. Role of transforming growth factor beta in human cancer. *J. Clin. Oncol.*, 23(9):2078–2093, 2005.

[37] M. R. Fabian, N. Sonenberg, and W. Filipowicz. Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.*, 79:351–379, 2010.

[38] D. Fange, O. G. Berg, P. Sjöberg, and J. Elf. Stochastic reaction-diffusion kinetics in the microscopic limit. *Proc. Natl. Acad. Sci. USA*, 107:19820–19825, 2010.

[39] E. C. Finger, R. S. Turley, M. Dong, T. How, T. A. Fields, and G. C. Blobe. TbetaRIII suppresses non-small cell lung cancer invasiveness and tumorigenicity. *Carcinogenesis*, 29(3):528–535, 2008.

[40] R. A. Fisher. *Statistical Methods, Experimental Design, and Statistical Inference*. Oxford University Press, Oxford, 1990.

[41] S. Frey, R. P. Richter, and D. Görlich. FG-rich repeats of nuclear pore proteins form a three-dimensional meshwork with hydrogel-like properties. *Science*, 314:815–7, 2006.

[42] C. Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer-Verlag, 2010.

[43] V. A. Gennarino, G. D'Angelo, G. Dharmalingam, S. Fernandez, G. Russolillo, R. Sanges, M. Mutarelli, V. Belcastro, A. Ballabio, P. Verde, M. Sardiello, and S. Banfi. Identification of microRNA-regulated gene networks by expression analysis of target genes. *Genome Res.*, 22(6):1163–1172, 2012.

[44] V. A. Gennarino, M. Sardiello, M. Mutarelli, G. Dharmalingam, V. Maselli, G. Lago, and S. Banfi. HOCTAR database: a unique resource for microRNA target prediction. *Gene*, 480(1-2):51–58, 2011.

[45] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2340–2361, 1977.

[46] D. T. Gillespie, A. Hellander, and L. R. Petzold. Perspective: Stochastic algorithms for chemical kinetics. *J. Chem. Phys.*, 138(170901), 2013.

[47] K. J. Gordon, M. Dong, E. M. Chislock, T. A. Fields, and G. C. Blobe. Loss of type III transforming growth factor beta receptor expression increases motility and invasiveness associated with epithelial to mesenchymal transition during pancreatic cancer progression. *Carcinogenesis*, 29(2):252–262, 2008.

[48] P. A. Gregory, A. G. Bert, E. L. Paterson, S. C. Barry, A. Tsykin, G. Farshid, M. A. Vadas, Y. Khew-Goodall, and G. J. Goodall. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat. Cell. Biol.*, 10(5):593–601, 2008.

[49] S. Gupta, S. E. Ellis, F. N. Ashar, A. Moes, J. S. Bader, J. Zhan, A. B. West, and D. E. Arking. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature Comm.*, 5, 2014.

[50] J. P. Hagan, E. Piskounova, and R. I. Gregory. Lin28 recruits the tutase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.*, 16:1021–1025, 2009.

[51] M. Hammell. Computational methods to identify miRNA targets. *Semin. Cell Dev. Biol.*, 21(7):738–744, 2010.

[52] E. M. Heinrich, J. Wagner, M. Krüger, D. John, S. Uchida, J. E. Weigand, B. Suess,

and S. Dimmeler. Regulation of miR-17-92a cluster processing by the microRNA binding protein SND1. *FEBS Lett.*, 587(15):2405–2411, 2013.

[53] S. Hellander, A. Hellander, and L. Petzold. Reaction-diffusion master equation in the microscopic limit. *Phys. Rev. E*, 85(042901), 2012.

[54] V. Helms. *Principles of Computational Cell Biology:Fluorescence Resonance Energy Transfer*. Weinheim: Wiley-VCH, 2008.

[55] I. Heo, C. Joo, Y. K. Kim, M. Ha, M. J. Yoon, J. Cho, K. H. Yeom, J. Han, and V. N. Kim. TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. *Cell*, 138:696–708, 2009.

[56] A. S. Hoffman. Hydrogels for biomedical applications. *Adv. Drug Deliv. Rev.*, 54:3–12, 2002.

[57] C.-W. Hsu, H.-F. Juan, and H.-C. Huang. Characterization of microRNA-regulated protein-protein interaction network. *Proteomics*, 8(10):1975–1979, 2008.

[58] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu, C.-H. Chien, M.-C. Wu, C.-Y. Huang, A.-P. Tsou, and H.-D. Huang. miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, 39:D163–D169, 2011.

[59] Y. W. A. Huang, C. R. Ruiz, E. C. Eyler, K. Lin, and M. K. Meffert. Dual regulation

of miRNA biogenesis generates target specificity in neurotrophin-induced protein synthesis. *Cell*, 148(5):933–946, 2012.

[60] K. M. Huber, M. S. Kayser, and M. F. Bear. Role for rapid dendritic protein synthesis in hippocampal mglur-dependent long-term depression. *Science*, 288:1254–1257, 2000.

[61] A. A. Hyman and K. Simons. Beyond oil and water–phase transitions in cells. *Science*, 337:1047–9, 2012.

[62] M. Inui, G. Martello, and S. Piccolo. MicroRNA control of signal transduction. *Nat. Rev. Mol. Cell Biol.*, 11(4):252–263, 2010.

[63] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T .P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[64] S. A. Isaacson. Relationship between the reaction-diffusion master equation and particle tracking models. *J. Phys. A: Math. Theor.*, 41(065003), 2008.

[65] S. A. Isaacson. The reaction-diffusion master equation as an asymptotic approximation of diffusion to a small target. *SIAM J. Appl. Math.*, 70:77–111, 2009.

[66] S. A. Isaacson. A convergent reaction-diffusion master equation. *J. Chem. Phys.*, 139:054–101, 2013.

[67] S. A. Isaacson and C. S. Peskin. Incorporating diffusion in compex geometries into stochastic chemical kinetics simulations. *SIAM J. Sci. Comput.*, 28:47–74, 2006.

[68] T. Iwasaki and Y. L. Wang. Cytoplasmic force gradient in migrating adhesive cells. *Biophys. J.*, 94:L35–7, 2008.

[69] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, 2000.

[70] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40(1):D109–D114, 2012.

[71] A. L. Kasinski and F. J. Slack. Epihenetics and genetics. MicroRNAs en route to the clinic: progress in validating and targeting microRNAs for cancer therapy. *Nat. Rev. Cancer*, 11(12):849–864, 2011.

[72] J. Keizer. Nonequilibrium statistical thermodynamics and the effect of diffusion on chemical reaction rates. *J. Phys. Chem.*, 86:5052–5067, 1982.

[73] A. Khademhosseini and R. Langer. Microengineered hydrogels for tissue engineering. *Biomaterials*, 28:5087–92, 2007.

[74] R. Kimmich. *Principles of Soft-Matter Dynamics: Basic Theories, Non-invasive Methods, Mesoscopic Aspects*. Springer Science & Business Media, Dordrecht, 2012.

[75] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–10, 2002.

[76] A. Kossenkov, F. J. Manion, E. Korotkov, T. D. Moloshok, and M. F. Ochs. ASAP: automated sequence annotation pipeline for web-based updating of sequence information with a local dynamic database. *Bioinformatics*, 19(5):675–676, 2003.

[77] J. Kost and M. McDermott. Combining dependent *P*-values. *Stat. Probabil. Lett.*, 60(2):183–190, 2002.

[78] M. Kumar, Z. Lu, A. A. Takwi, W. Chen, N. S. Callander, K.S. Ramos, K. H. Young, and Y. Li. Negative regulation of the tumor suppressor p53 gene by microRNAs. *Oncogene*, 30(7):843–853, 2011.

[79] C. Kumar-Sinha, S. A. Tomlins, and A. M. Chinnaiyan. Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer*, 8:497–511, 2008.

[80] S. K. Lai, Y. Y. Wang, D. Wirtz, and J. Hanes. Micro- and macrorheology of mucus. *Adv. Drug Deliv. Rev.*, 61:86–100, 2009.

[81] A. Lal, F. Navarro, C. A. Maher, L. E. Maliszewski, N. Yan, E. O'Day, D. Chowdhury, D.M. Dykxhoorn, P. Tsai, O. Hofmann, K.G. Becker, M. Gorospe, W. Hide, and J. Lieberman. miR-24 inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements. *Mol. Cell*, 11(3):610–625, 2009.

[82] P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(559), 2008.

[83] T. D. Le, L. Liu, B. Liu, A. Tsykin, G. J. Goodall, K. Satou, and J. Li. Inferring microRNA and transcription factor regulatory networks in heterogeneous data. *BMC Bioinformatics*, 14:92, 2013.

[84] A. Lee, H. Nakamura, Lin Y. C. Afshar, A. S., M. Tanigawa, A. Suarez, S. Razavi, J. M. McCaffery, R. DeRose, D. Bobb, W. Hong, S. B. Gabelli, J. Goutsias, and T. Inoue. Intracellular production of synthetic RNA granules by ligand-yielded multivalent enhancers. *Under Revision*, 2016.

[85] J. Y. Lee, J. Colinas, J. Y. Wang, D. Mace, U. Ohler, and P. N. Benfey. Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots. *Proc. Natl. Acad. Sci. USA*, 103(15):6055–6060, 2006.

[86] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, 11:733–739, 2010.

[87] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, 3(9):e161, 2007.

[88] J. Li, X. Hua, M. Haubrock, J. Wang, and E. Wingender. The architecture of the gene regulatory networks of different tissues. *Bioinformatics*, 28(18):i509–i514, 2012.

[89] P. Li, S. Banjade, H. C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q. X. Jiang, B. T. Nixon, and M. K. Rosen. Phase transitions in the assembly of multivalent signaling proteins. *Nature*, 483:336–341, 2012.

[90] Y. Li, D. Kong, A. Ahmad, B. Bao, G. Dyson, and F.H. Sarkar. Epigenetic deregulation of miR-29a and miR-1256 by isoflavone contributes to the inhibition of prostate cancer cell growth and invasion. *Epigenetics*, 7(8):940–949, 2012.

[91] Y. R. Li, O. D. King, J. Shorter, and A. D. Gitler. Stress granules as crucibles of als pathogenesis. *J. Cell Biol.*, 201:361–72, 2013.

[92] H. Liang and W.-H. Li. MicroRNA regulation of human protein protein interaction network. *RNA*, 13(9):1402–1408, 2007.

[93] O. Lieleg and K. Ribbeck. Biological hydrogels as selective diffusion barriers. *Trends Cell Biol.*, 21:543–551, 2011.

[94] J. Lipková, K. C. Zygalakis, S. J. Chapman, and R. Erban. Analysis of Brownian dynamics simulations of reversible bimolecular reactions. *SIAM J. Appl. Math.*, 71:714–730, 2011.

[95] Y. Liu, B. Yin, C. Zhang, L. Zhou, and J. Fan. Hsa-let-7a functions as a tumor suppressor in renal cell carcinoma cell lines by targeting c-myc. *Biochem. Biophys. Res. Commun.*, 417(1):371–375, 2012.

[96] P. Lopez-Serra and M. Esteller. DNA methylation-associated silencing of tumor-suppressor microRNAs in cancer. *Oncogene*, 31(13):1609–1622, 2012.

[97] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz, and T. R. Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435:834–838, 2005.

[98] A. Lujambio, G. A. Calin, A. Villanueva, S. Ropero, M. Sánchez-Cspedes, D. Blanco, L. M. Montuenga, S. Rossi, M. S. Nicoloso, W. J. Faller, W. M. Gallagher, S. A. Eccles, C. M. Croce, and M. Esteller. Genetic unmasking of an epigenetically silenced microRNA in human cancer cells. *Proc. Natl. Acad. Sci. USA*, 105(36):13556–13561, 2008.

[99] A. Lujambio, S. Ropero, E. Ballestar, M. F. Fraga, C. Cerrato, F. Setién, S. Casado, A. Suarez-Gauthier, M. Sanchez-Cespedes, A. Git, I. Spiteri, P. P. Das, C. Caldas, E. Miska, and M. Esteller. Genetic unmasking of an epigenetically silenced microRNA in human cancer cells. *Cancer Res.*, 67(4):1424–1429, 2007.

[100] D. Medici, E. D. Hay, and B. R. Olsen. Snail and Slug promote epithelial-mesenchymal transition through beta-catenint-cell factor-4-dependent expression of transforming growth factor-beta3. *Mol. Biol. Cell.*, 19(11):4875–4887, 2008.

[101] P. P. Medina, M. Nolde, and F. J. Slack. OncomiR addiction in an in vivo model of microRNA-21-induced pre-B-cell lymphoma. *Nature*, 467(7311):86–90, 2010.

[102] J. T. Mendell. MiRiad roles for the miR-17-92 cluster in development and disease. *Cell*, 133(2):217–222, 2008.

[103] K. Miyazono. Transforming growth factor-beta signaling in epithelial-mesenchymal transition and progression of cancer. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.*, 85(8):314–323, 2009.

[104] E. Mogilyansky and I. Rigoutsos. The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell Death Differ.*, 20(12):1603–1614, 2013.

[105] K. A. Mosiewicz, L. Kolb, A. J. Van Der Vlies, M. M. Martino, P. S. Lienemann, J. A. Hubbell, M. Ehrbar, and M. P. Lutolf. In situ cell manipulation through enzymatic hydrogel photopatterning. *Nat. Mater.*, 12:1072–8, 2013.

[106] W. Mulyasasmita, J. S. Lee, and S. C. Heilshorn. Molecular-level engineering of protein physical hydrogels for predictive sol-gel phase behavior. *Biomacromolecules*, 12:3406–11, 2011.

[107] S. Obad, C. O. dos Santos, A. Petri, M. Heidenblad, O. Broom, C. Ruse, C. Fu, M. Lindow, J. Stenvang, E. M. Straarup, H. F. Hansen, T. Koch, D. Pappin, G. J. Hannon, and S. Kauppinen. OncomiR addiction in an in vivo model of microRNA-21-induced pre-B-cell lymphoma. *Nat. Genet.*, 43(4):371–378, 2011.

[108] K. A. O'Donnell, E. A. Wentzel, K. I. Zeller, C. V. Dang, and J. T. Mendell. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, 435:839–843, 2005.

[109] V. Olive, Q. Li, and L. He. mir-17-92: a polycistronic oncomir with pleiotropic functions. *Immunol. Rev.*, 253(1):158–166, 2013.

[110] H. Osada and T. Takahashi. let-7 and mir-17-92: small-sized major players in lung cancer development. *Cancer Sci.*, 102(1):9–17, 2011.

[111] M. Ozen, C. J. Creighton, M. Ozdemir, and M. Ittmann. Widespread deregulation of microRNA expression in human prostate cancer. *Oncogene*, 27(12):1788–1793, 2008.

[112] S. Y. Park, M. S. Jeong, and S. B. Jang. In vitro binding properties of tumor suppressor p53 with PUMA and NOXAs. *Biochem. Biophys. Res. Commun.*, 420(2):350–356, 2012.

[113] Z. Paroo, X. Ye, S. Chen, and Q. Liu. Phosphorylation of the human microRNA-generating complex mediates MAPK/Erk signaling. *Cell*, 139:112–122, 2009.

[114] A. E. Pasquinelli. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat. Rev. Genet.*, 13(4):271–282, 2012.

[115] F. Petrocca, R. Visone, M. R. Onelli, M. H. Shah, M. S. Nicoloso, I. de Martino, D. Iliopoulos, E. Pilozzi, C. G. Liu, M. Negrini, L. Cavazzini, S. Volinia, H. Alder,

L. P. Ruco, G. Baldassarre, C. M. Croce, and A. Vecchione. E2F1-regulated microRNAs impair TGFbeta-dependent cell-cycle arrest and apoptosis in gastric cancer. *Cancer Cell*, 13(3):272–286, 2008.

[116] J. L. Pfaltz. Chordless cycles in networks. *Proceedings of the 29th IEEE International Conference on Data Engineering Workshops (ICDEW), Brisbane, Australia, April 8-12*, pages 223–228, 2013.

[117] K. Polyak and R. A. Weinberg. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nature Rev. Cancer*, 9:265–273, 2009.

[118] K. P. Porkka, M. J. Pfeiffer, K. K. Waltering, R. L. Vessella, T. L. J. Tammela, and T. Visakorpi. MicroRNA expression profiling in prostate cancer. *Cancer Res.*, 67:6130–6135, 2007.

[119] A. Re, D. Corá, D. Taverna, and M. Caselle. Genome-wide survey of microRNA transcription factor feed-forward regulatory circuits in human. *Mol. Biosyst.*, 5(8):854–867, 2009.

[120] I. Rhee, K. E. Bachman, B. H. Park, K. W. Jair, R. W. C. Yen, K. E. Schuebel, H. Cui, A. P. Feinberg, C. Lengauer, K. W. Kinzler, S. B. Baylin, and B. Vogelstein. DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature*, 416(6880):552–556, 2002.

[121] J. V. Rodrguez, J. A. Kaandorp, M. Dobrzyński, and J. G. Blom. Spatial stochastic modelling of the phosphoenolpyruvate-dependent phosphotransferase (PTS) pathway in escherichia coli. *Bioinformatics*, 22:1895–1901, 2006.

[122] M. Rodríguez-Paredes and M. Esteller. Cancer epigenetics reaches mainstream oncology. *Nat. Med.*, 17(3):330–339, 2011.

[123] K. M. Roy. *Sulfones and Sulfoxides: Ullmann's Encyclopedia of Industrial Chemistry*. Weinheim: Wiley-VCH, 2002.

[124] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and H. W. Mewes. CORUM: the comprehensive resource of mammalian protein complexes2009. *Nucleic Acids Res.*, D497-D501, 2009.

[125] M. Sachdeva, S. Zhu, F. Wu, H. Wu, V. Walia, S. Kumar, R. Elble, K. Watabe, and Y.-Y. Mo. p53 represses c-Myc through induction of the tumor suppressor mir-145. *Proc. Natl. Acad. Sci. USA*, 106(9):3207–3012, 2009.

[126] T. Saito and P. Sætrom. MicroRNAs-targeting and target prediction. *Nat. Biotechnol.*, 27(3):243–249, 2010.

[127] Y. Saito, G. Liang, G. Egger, J. M. Friedman, J. C. Chuang, G. A. Coetzee, and P. A. Jones. Specific activation of microRNA-127 with downregulation of the proto-oncogene BCL6 by chromatin-modifying drugs in human cancer cells. *Cancer Cell*, 9(6):435–443, 2006.

[128] S. Sass, S. Dietmann, U. C. Burk, S. Brabletz, D. Lutter, A. Kowarsch, K. F. Mayer, T. Brabletz, A. Ruepp, F. J. Theis, and Y. Wang. MicroRNAs coordinately regulate protein complexes. *BMC Syst. Biol.*, 5:136, 2011.

[129] A. Schaefer, M. Jung, G. Kristiansen, M. Lein, M. Schrader, K. Miller, C. Stephan, and K. Jung. MicroRNAs and cancer: current state and future perspectives in urologic oncology. *Urol. Oncol.*, 28(1):4–13, 2008.

[130] G. M. Schratt, E. A. Nigh, W. G. Chen, L. Hu, and M. E. Greenberg. BDNF regulates the translation of a select group of mRNAs by a mammalian target of rapamycin-phosphatidylinositol 3-kinase-dependent pathway during neuronal development. *J. of Neurosci.*, 24(33):7366–7377, 2004.

[131] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 2008.

[132] M. Setty, K. Helmy, A. A. Khan, J. Silber, A. Arvey, F. Neezen, P. Agius, J. T. Huse, E. C. Holland, and C. S. Leslie. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol. Syst. Biol.*, 8:605, 2012.

[133] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel. Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Comput. Biol.*, 3(7):e131, 2007.

[134] J. Shi and M. G. Walker. Gene set enrichment analysis (GSEA) for interpreting gene expression profiles. *Curr. Bioinform.*, 2(2):133–137, 2007.

[135] X. B. Shi, L. Xue, A. H. Ma, C. G. Tepper, H. J. Kung, and R. W. White. miR-125b promotes growth of prostate cancer xenograft tumor through targeting pro-apoptotic genes. *Prostate*, 71(5):538–549, 2011.

[136] K. Sikand, S. D. Slane, and G. C. Shukla. Intrinsic expression of host genes and intronic miRNAs in prostate carcinoma cells. *Cancer Cell Int.*, 9:21, 2009.

[137] G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3(1):3, 2004.

[138] M. W. Sneddon, J. R. Faeder, and T. Emonet. Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. *Nat. Methods*, 8:177–183, 2011.

[139] W. Speir and M. F. Ochs. Updating annotations with the distributed annotation system and the automated sequence annotation pipeline. *Bioinformatics*, 28(21):2858–2859, 2012.

[140] A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. Animal microRNAs confer robustness to gene expression and have a significant impact on 3' UTR evolution. *Cell*, 123(6):1133–1146, 2005.

[141] N. Su, Y. Wang, M. Qian, and M. Deng. Combinatorial regulation of transcription factors and microRNAs. *BMC Syst. Biol.*, 4:150, 2010.

[142] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, 102(43):15545–15550, 2005.

[143] P. Sumazin, X. Yang, H. S. Chiu, W. J. Chung, A. Iyer, D. Llobet-Navas, P. Rajbhandari, M. Bansal, P. Guarnieri, J. Silva, and A. Califano. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell*, 147(2):370–381, 2011.

[144] C. Sun, A. Dobi, A. Mohamed, H. Li, R. L. Thangapazham, B. Furusato, S. Shaheduzzaman, S. H. Tan, G. Vaidyanathan, E. Whitman, D. J. Hawksworth, Y. Chen, M. Nau, V. Patel, M. Vahey, J. S. Gutkind, T. Sreenath, G. Petrovics, I. A. Sesterhenn, D. McLeod, and S. G. Srivastava. TMPRSS2-ERG fusion, a common genomic alteration in prostate cancer activates C-MYC and abrogates prostate epithelial differentiation. *Oncogene*, 27(40):5348–5353, 2008.

[145] J. Sun, X. Gong, B. Purow, and Z. Zhao. Uncovering microRNA and transcription factor mediated regulatory networks in glioblastoma. *PLoS Comput. Biol.*, 8(7):e1002488, 2012.

[146] Y. M. Sun, K. Y. Lin, and Y. Q. Chen. Diverse functions of miR-125 family in different cell contexts. *J. Hematol. Oncol.*, 6:6, 2013.

[147] J. Szczyrba, E. Löprich, S. Wach, V. Jung, G. Unteregger, S. Barth, R. Grobholz, W. Wieland, R. Stöhr, A. Hartmann, B. Wullich, and F. Grässer. The microRNA profile of prostate carcinoma obtained by deep sequencing. *Mol. Cancer Res.*, 8(4):529–538, 2010.

[148] K. Takahashi, S. N. Arjunan, and M. Tomita. Space in systems biology of signaling pathways-towards intracellular molecular crowding in silico. *FEBS Lett.*, 579:1783–1788, 2005.

[149] B. Tang, T. Wang, H. Wan, L. Han, X. Qin, Y. Zhang, J. Wang, C. Yu, F. Berton, W. Francesconi, and J.R. Yates. Fmr1 deficiency promotes age-dependent alterations in the cortical synaptic proteome. *Proc. Natl. Acad. Sci., USA*, 112(34):E4697–E4706, 2015.

[150] B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. S. Carver, V. K. Arora, P. Kaushik, E. Cerami, B. Reva, Y. Antipin, N. Mitsiades, T. Landers, I. Dolgalev, J. E. Major, M. Wilson, N. D. Socci, A. E. Lash, A. Heguy, Eastham J. A., Scher H. I., V. E. Reuter, P. T. Scardino, C. Sander, C. L. Sawyers, and W. L. Gerald. Integrative genomic profiling of human prostate cancer. *Cancer Cell*, 18(1):11–22, 2010.

[151] S. A. Tomlins, B. Laxman, S. Varambally, X. Cao, J. Yu, B. E. Helgeson, Q. Cao,

J. R. Prensner, M. A. Rubin, R. B. Shah, R. Mehra, and A. M. Chinnaiyan. Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia*, 10(2):177–188, 2008.

[152] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310(5748):644–648, 2005.

[153] A. W. Tong, P. Fulgham, C. Jay, P. Chen, I. Khalil, S. Liu, N. Senzer, A. C. Eklund, J. Han, and J. Nemunaitis. MicroRNA profile analysis of human prostate cancers. *Cancer Gene Ther.*, 16(3):206–216, 2009.

[154] D. H. Tran, K. Satou, T. B. Ho, and T. H. Pham. Computational discovery of miR-TF regulatory modules in human genome. *Bioinformation*, 4(8):371–377, 2010.

[155] J. Tsang, J. Zhu, and A. van Oudenaarden. MicroRNA-mediated feedback and feed-forward loops are recurrent network motifs in mammals. *Mol. Cell*, 26(5):753–767, 2007.

[156] J. S. Tsang, M. S. Ebert, and A. van Oudenaarden. Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Mol. Cell*, 38(1):140–153, 2010.

[157] R. S. Turley, E. C. Finger, N. Hempel, T. How, T. A. Fields, and G. C. Blobe. The

type III transforming growth factor-beta receptor as a novel tumor suppressor gene in prostate cancer. *Cancer Res.*, 67(3):1090–1098, 2007.

[158] N. G. Van Kampen. Stochastic processes in physics and chemistry. *North-Holland*, 3rd ed., 2007.

[159] A. Ventura, A. G. Young, M. M. Winslow, L. Linault, A. Meissner, S. J. Erkeland, J. Newman, R. T. Bronson, D. Crowley, J. R. Stone, R. Jaenisch, P. A. Sharp, and T. Jacks. Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. *Cell*, 132(5):875–886, 2008.

[160] I. S. Vlachos and A. G. Hatzigeorgiou. Online resources for miRNA analysis. *Clin. Biochem.*, 46(10-11):879–900, 2013.

[161] S. Volinia, G. A. Calin, C.-G. Liu, S. Ambs, A. Cimmino, F. Petrocca, R. Visone, M. Iorio, C. Roldo, M. Ferracin, R. L. Prueitt, N. Yanaihara, G. Lanza, A. Scarpa, A. Vecchione, M. Negrini, C. C. Harris, and C. M. Croce. A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl. Acad. Sci. USA*, 103(7):2257–2261, 2006.

[162] S. Wach, E. Nolte, J. Szczyrba, R. Stöhr, A. Hartmann, T. Ørntoft, L. Dyrskjøt, E. Eltze, W. Wieland, B. Keck, A. B. Ekici, F. Grässer, and B. Wullich. MicroRNA profiles of prostate carcinoma detected by multiplatform microRNA screening. *Int. J. Cancer*, 130(3):611–621, 2012.

[163] J. Wang, M. Lu, C. Qiu, and Q. Cui. TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res.*, 38:D119–D122, 2010.

[164] L. Wang, P. Jia, R. D. Wolfinger, X. Chen, and Z. Zhao. Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics*, 98(1):1–8, 2011.

[165] S. C. Weber and C. P. Brangwynne. Getting RNA and protein in phase. *Cell*, 149:1188–1191, 2012.

[166] K. Weis. The nuclear pore complex: oily spaghetti or gummy bear? *Cell*, 130:405–407, 2007.

[167] C. Wels, S. Joshi, P. Koefinger, H. Bergler, and H. Schaider. Transcriptional activation of ZEB1 by Slug leads to cooperative regulation of the epithelial-mesenchymal transition-like phenotype in melanoma. *J. Invest. Dermatol.*, 131(9):1877–1885, 2011.

[168] D. B. West. *Introduction to Graph Theory*. Prentice Hall, 2000.

[169] M. Westberg. Combining independent statistical tests. *Statistician*, 34:287–296, 1985.

[170] E. Wingender. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.*, 9(4):326–332, 2008.

[171] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, 37:D105–D110, 2009.

[172] J. Xu, S. Lamouille, and R. Derynck. TGF-β-induced epithelial to mesenchymal transition. *Cell Res.*, 19(1):156–172, 2009.

[173] J. Xu and C. Wong. A computational screen for mouse signaling pathways targeted by microRNA clusters. *RNA*, 14(7):1276–1283, 2008.

[174] Z. Yan, P. K. Shah, S. B. Amin, M. K. Samur, N. Huang, X. Wang, V. Misra, H. Ji, D. Gabuzda, and C. Li. Integrative analysis of gene and miRNA expression profiles with transcription factor-miRNA feed-forward loops identifies regulators in human cancers. *Nucleic Acids Res.*, 40(17):e135, 2012.

[175] J. Yang, M. I. Monine, J. R. Faeder, and W. S. Hlavacek. Kinetic Monte Carlo method for rule-based modeling of biochemical networks. *Phys. Rev. E*, 78(031910), 2008.

[176] H. Yu, K. Tu, Y. J. Wang, J. Z. Mao, L. Xie, Y. Y. Li, and Y. X. Li. Combinatorial network of transcriptional regulation and microRNA regulation in human cancer. *BMC Syst. Biol.*, 6:61, 2012.

[177] B. Zacher, K. Abnaof, S. Gade, E. Younesi, A. Tresch, and H. Fröhlich. Joint Bayesian inference of condition specific miRNA and transcription factor activities

from combined gene and microRNA expression data. *Bioinformatics*, 28(13):1714–1720, 2012.

[178] Y. Zhou, J. Ferguson, J. T. Chang, and Y. Kluger. Inter- and intra-combinatorial regulation by transcription factors and microRNAs. *BMC Genomics*, 8:396, 2007.

Vita

Ali received his B.Sc. and M.Sc. (with honors) in Electrical Engineering respectively from Shiraz University and Sharif University of Technology, the top engineering school in Iran. At JHU, together with his PhD studies, he also received his M.S.E. in Applied Mathematics and Statistics as a joint a program. As a PhD student, Ali initiated and led three independent research works, all of which shared the focus of translational research in systems biology. His research in Cancer Systems Biology was on the identification of MicroRNA-mediated gene regulatory networks in cancer. MicroRNAs, as a relatively new class of biological molecules, are broadly implicated in fundamental gene regulation, and they offer great prospects for cancer therapy due to their diagnostic as well as therapeutic potential. His second work involved an interdisciplinary and challenging research on synthetically developing and characterizing a biomaterial that would act as a molecular sieve to control the passage of biomolecules in living cells. The research has the potential to offer a novel, universal way of analyzing

and manipulating biochemical and biophysical properties of cellular processes. The biocompatible nature of proteins involved presents them as an attractive choice for fabrication of smart, programmable, biopolymers which have a range of applications in drug delivery, materials science, and biotechnology. His third research area focused on how MicroRNA-mediated gene regulation affects the synaptic, neuronal and cognitive function in the brain. Their collaborative research identified dysregulation of particular MicroRNAs in Fragile X Syndrome, which represents the most common inherited form of intellectual disability. This work has led to the discovery of a novel, blood-based biomarker in an animal model, and their ongoing work aims at the development of this biomarker to be used as a diagnostic tool to detect diseases such as Fragile X Syndrome and ultimately a therapeutic approach for the treatment of the Autism Spectrum Disorders. Besides his scientific endeavors, Ali is the president and founder of Technology Entrepreneurship club which has been instrumental in organizing and coordinating several distinguished, technology-related events such as the IBM Bluemix Hackathon held at Johns Hopkins in 2014. He was the recipient of the prestigious Entrepreneurs Choice Award in the Mid-Atlantic Regional Finals of Venture Capital Investment Competition 2015. Ali was named as one of the 50 Leaders in the Mid-Atlantic region, in the Leaders of Tomorrow Summit 2015 hosted by AstraZeneca, with the goal to bring together the next generation of biotech leaders with established leaders and to ignite disruptive and innovative ideas and help grow the Mid-Atlantics future bioeconomy.